ADASS 2023 Tutorial Plan
Large Survey DataBase + HiPSCat

Samuel Wyatt[1], Mario Juric[1 2], Andrew Connolly[1], Melissa DeLucchi[3], Sean McGuire[3], Max West[1],
Rachel Mandelbaum[3], Jeremy Kubica[3], Carl Christofferson[1],
[1] University of Washington, Seattle, WA
[2] Rubin Construction Project, Seattle, WA
[3] Carnegie Mellon University, Pittsburgh, PA

Working title for the tutorial:

LSDB and HiPSCat: Joint Distributed Analysis of LSST-Scale Datasets

Description of the main topic:

The present decade will be marked by growth of large survey catalogs, both in their number and scale. Joint analysis of such catalogs has historically shown itself to be tremendously useful (e.g. enabling multi-wavelength or time-domain studies), with its importance likely to rise even further. Yet, with the increase in scale towards PBs of data, joint analysis – even at a catalog level – becomes a complex data management problem that few astronomers are equipped to tackle with present-day technology. Here we present HiPSCat, a format for efficient and queryable storage of large datasets, and LSDB (Large Survey DataBase), a Python framework that enables distributed cross-matching and analysis of astronomical datasets at LSST scale (O(10B) sources). The HiPSCat format - framework-independent and built as an extension of the well-known IVOA HiPS standard - provides intelligent (balanced) spatial partitioning and enables scalable serving of PB-scale datasets (via HTTP) using Parquet for efficient storage. The LSDB framework enables distributed computing and cross-matching on HiPSCat-formatted datasets. Leveraging broadly adopted community libraries such as astropy, Pandas, and Dask, LSDB presents a user-friendly API approachable to astronomers. The goal of LSDB is to enable the user to focus on the science aspects of their tasks, leaving the difficult data management aspects (distribution, resiliency) to the framework.

Primary learning objectives:

1. *How LSDB can achieve your large scale science:* Out of the box, LSDB comes equipped with the potential to perform various spatial analysis (like cone-searching, and cross-matching), along with time-series analysis (e.g. large scale lightcurve analysis). Not only will LSDB be equipped with these methods, but since it is built upon the `dask.dataframe` distributed framework, users will be able to define their own functions and map them across the catalogs. If a user wants to use their own cross-match algorithm, they can easily tie it into our HiPSCat framework through the `dask.dataframe` library, which we will provide adequate documentation on.

2. *The strengths of HiPSCat partitioning structure:* It enables storage of astronomical datasets in a way that equalizes the number of rows per partition, yet keeps spatially adjacent objects together. Once two catalogs are partitioned in the HiPSCat manner, distributed, joint-spatial analysis is trivial between them.

Detailed tutorial structure:

- *HiPSCat overview (10-15min)*: Here we will give a presentation on the HiPSCat concept: what it is, how we partition our catalogs, and how the schema enables distributed, joint analysis for extremely large datasets.
- *HiPSCat datasets and LSDB framework access (5min)*: We will provide users with access to our Jupyter Hub, and have them import the sample notebooks so that they may follow along during the tutorial. Here they will have access to the LSDB python framework, and have the ability to connect to the S3 buckets that hold the sample, pre-partitioned HiPSCat datasets.
- *Basic HiPSCat functionality (15-20min)*: We will demonstrate the key functionality behind the LSDB software package when it comes to loading catalogs, gathering metadata, and basic visualization. We will also demonstrate how to perform basic functionality on singular HiPSCats: Querying based on column parameters, performing cone-searches, and writing custom functions based on these results.
- *Joint analysis of many HiPSCats (15-20min)*: We will show the user how to perform multi-catalog functionality (joining and cross-matching) on HiPSCats. We will also demonstrate the power of *chaining* functionality, where we can perform querying, cross-matching, and running custom functions in one line of code in a distributed manner.
- *Time-Series analysis of HiPSCats (15-20min*: We will demonstrate the capability of performing time-series analysis with HiPSCats. We will show how we can join objects to their sources, and thusly generate light-curves from the results. We will also show examples of running custom functions on the light-curves; like SNe template fitting, variable star fitting, etc.
- *Importing your own HiPSCat datasets (10-15min)*: We will finally demonstrate how to convert your own datasets into HiPSCats with our import tool. The user will be able to generate their own object, source, and margin catalogs; allowing them to be able to perform LSDB functionality on their own datasets.

Total time: *1.25-1.5 hr*.

Tutorial material

Example Jupyter notebooks can be found at this github repo in the `nb/` folder: github.com/swy-att7/ADASS_LSDB_tutorial. There is also a link to short presentation that is intended for the *HiPSCat overview* tutorial section in the `pres/` folder

List of what the participant will need

- Users will only need a laptop that has internet access, and a github account we can invite to our Jupyter-hub.

We will provide access to a Jupyter-hub that users can log-into and have immediate access to the LSDB python framework and the sample datasets we will provide.