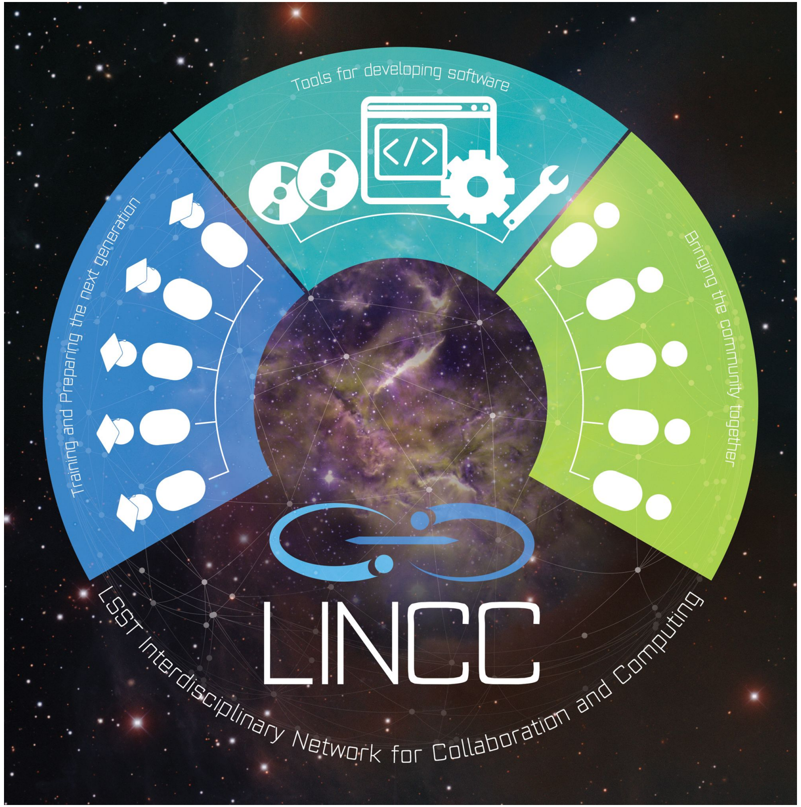
LSDB Overview

ADASS Tutorial
Samuel Wyatt + LINCC members
11/08/2023

LINCC

- **L**SST **I**nterdisciplinary **N**etwork for **C**ollaboration and **C**omputing
- Science Frameworks:
  - Scalable Spatial Analysis (**LSDB**)
  - Time Domain (TAPE & **LSDB**)
  - Scalable Faint Object Detection (KBMOD)
  - Comprehensive Photo-Z infrastructure (RAIL)

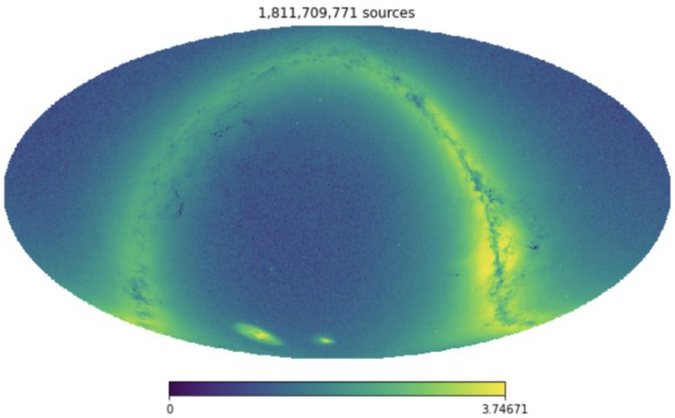

Tools for developing software
Training and Preparing the next generation
Bringing the community together
LINCC
LSST Interdisciplinary Network for Collaboration and Computing

LSDB

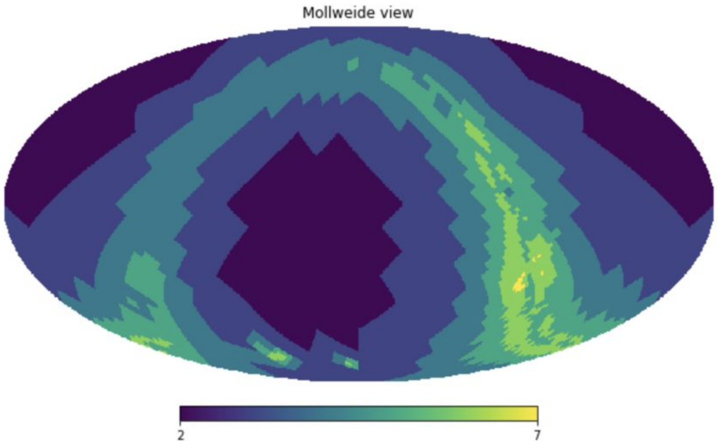
- **L**arge **S**urvey **D**ata**B**ase
- Supporting LSST science questions requires key functionality in an analysis framework with the ability to:
    - Store and manipulate catalog data at scale
    - Perform distributed computation over this data
    - Use spatial structure within searches and statistical computation
    - Interoperate with data from other surveys
    - Access these catalogs without having to directly download them.

LSDB: HiPSCATs

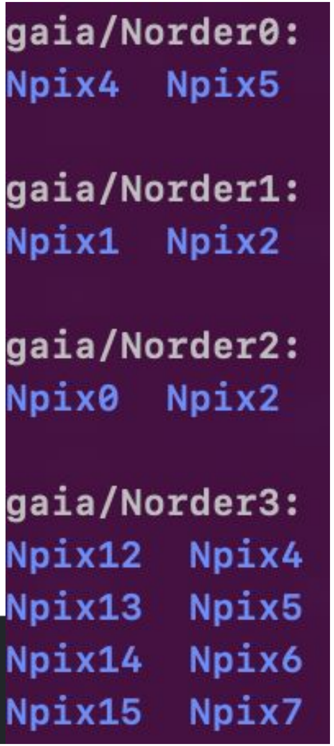
- Partition the source catalogs in a way to enable efficient/scalable analysis.
  - Input large source catalog (list of files csv/parquet/fits)
  - Index the sources in healpix space based on catalog density per index



Gaia DR2 Catalog Counts (log scale)



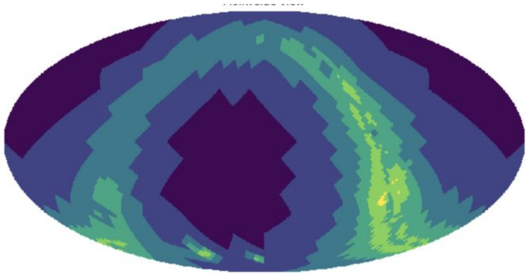
Visualization of file storage (color = healpix level)
3933 partitions of similar size (128-256 MB)

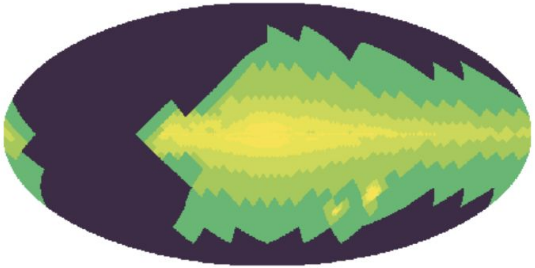


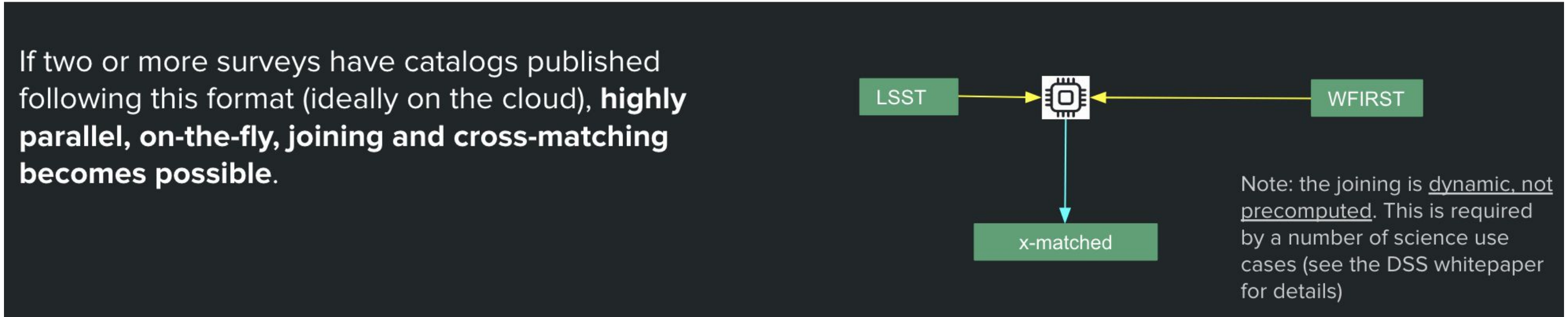
```
gaia/Norder0:
Npix4   Npix5

gaia/Norder1:
Npix1   Npix2

gaia/Norder2:
Npix0   Npix2

gaia/Norder3:
Npix12   Npix4
Npix13   Npix5
Npix14   Npix6
Npix15   Npix7
```

LSDB: HiPSCATs

- Once two catalogs are in this format parallelized spatial analysis is (should be) trivial and fast







If two or more surveys have catalogs published following this format (ideally on the cloud), **highly parallel, on-the-fly, joining and cross-matching becomes possible**.

LSST

WFIRST

x-matched

Note: the joining is dynamic, not precomputed. This is required by a number of science use cases (see the DSS whitepaper for details)

LSDB: Spatial Analysis Requirments

- Use-case requirements: Real-Time and Offline Static
  - The real-time component would entail low-latency matching of O(10k) sources to O(10) catalogs each holding O(1Bn) sources.
    - E.G. matching a LSST single image to multiple catalogs
  - The static component would need to support matching of O(10Bn) x O(1Bn) object catalog.
    - E.G. Matching the full LSST source catalog to GAIA's source catalog.
  - Retain general spatial querying (easy with healpy):
    - Cone searches (objects in radius of RA, DEC)
    - Polygon
- Technical Requirements:
  - Framework with scalable distributed processing: Dask, Ray, PySpark
    - Prototyping with Dask currently
  - Friendly user-interface: Importable python libraries, command line interfaces
  - Large data hosting for commonly used source catalogs already partitioned
    - (lengths ~1Bn)

LSDB - Time Series

- HiPSCat Association Tables (in development)
    - Joining objects to their individual observations (sources) and retaining that relation for easy of querying.
    - On the fly access to light curves
    - Can also be applied to precomputed cross-matches

LSDB - use cases

- Chaining methods:
  - Filtering
  - Spatial Querying
  - Cross-matching
- Time-series analysis
- Applying custom functions
- Real-time and Static components

Contributed usecases

LSDB

- Tutorial Notebooks
  - ADASS Tutorial