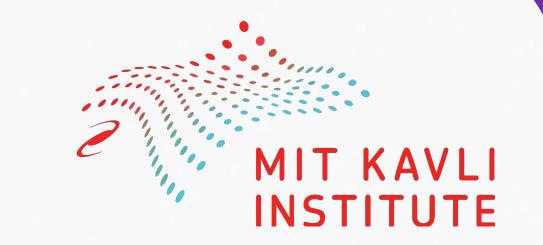# The Chandra Source Catalog 2.0: Building The Catalog
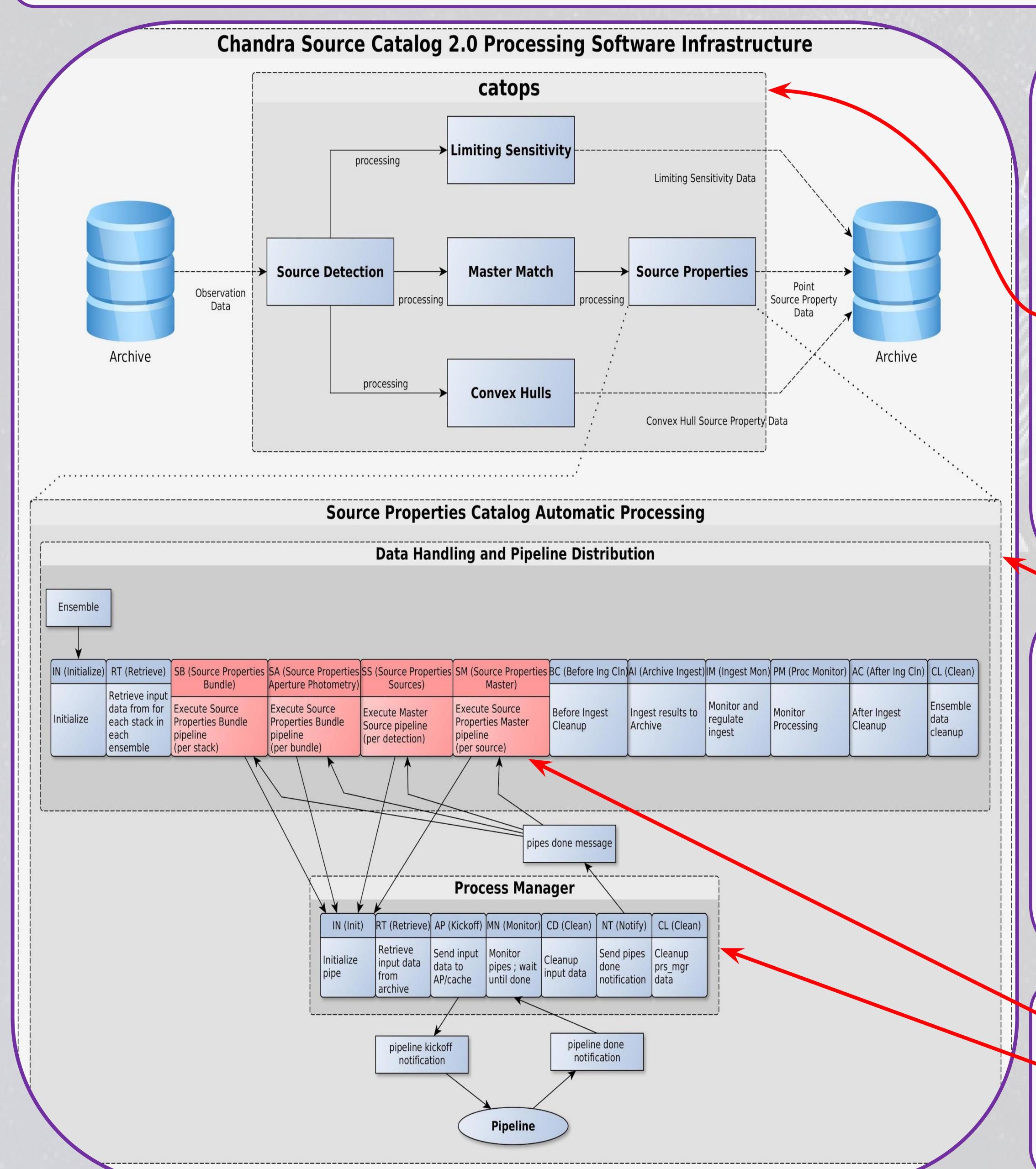
John D. Grier[1], David A. Plummer[1], Christopher Allen[1], Craig S. Anderson[1], Jamie A. Budynkiewicz[1], Douglas Burke[1], Judy C. Chen[1], Francesca Civano[1], Raffaele D'Abrusco[1], Stephen M. Doe[2], Ian N. Evans[1], Janet D. Evans[1], Giuseppina Fabbiano[1], Danny G. Gibbs II[1], Kenny J. Glotfelty[1], Dale E. Graessle[1], Roger M. Hain[1], Diane M. Hall[3], Peter N. Harbo[1], John C. Houck[1], Jennifer Lauer[1], Omar Laurino[1], Nicholas Lee[1], J. Rafael Martinez-Galarza[1], Michael L. McCollough[1], Jonathan C. McDowell[1], Warren McLaughlin[1], Joseph B. Miller[1], Douglas L. Morgan[1], Amy E. Mossman[1], Dan T. Nguyen[1], Joy S. Nichols[1], Michael A. Nowak[4], Charles Paxson[1], Francis A. Primini[1], Arnold H. Rots[1], Aneta Siemiginowska[1], Beth A. Sundheim[1], Michael S. Tibbetts[1], David W. Van Stone[1], Panagoula Zografou[1]

[1]Smithsonian Astrophysical Observatory [2]formerly Smithsonian Astrophysical Observatory [3]Northrop Grumman Mission Systems
[4]MIT Kavli Institute for Astrophysics and Space Research

Abstract

To build release 2.0 of the Chandra Source Catalog (CSC 2.0), we require scientific software tools and processing pipelines to evaluate and analyze the data. Additionally, software and hardware infrastructure is needed to coordinate and distribute pipeline execution, manage data i/o, and handle data for Quality Assurance (QA) intervention. We also provide data product staging for archive ingestion.

Release 2.0 utilizes a database driven system for operations and production. Included are five instances of the Automatic Processing (AP) system (Source Detection, Master Match, Source Properties, Limiting Sensitivity and Convex Hulls) and a High-Performance Computing Cluster (HPCC) that is managed to provide efficient catalog processing. In this poster we highlight the internal systems developed to meet the CSC 2.0 challenge.

Chandra Source Catalog 2.0 Processing Software Infrastructure

catops

Source Properties Catalog Automatic Processing

Data Handling and Pipeline Distribution

Process Manager

Operations and Production

- Release 2.0 of the Chandra Source Catalog (CSC 2.0) is built using an infrastructure of hardware and software managed by a database and executed on an High-Performance Computing Cluster (HPCC). This infrastructure:
  - Executes the data analysis tools that are organized into pipelines that process the data in order to determine the source positions and properties that populate the catalog.
  - Produces the tabulated properties for the sources along with corresponding file-based data products (in FITS format) which are stored in the archive and accessible to the user.
- An sqlite database called "catops" is used to manage processing. This database:
  - Defines the five "CSC Modules" that make up the CSC 2.0 processing system: Source Detection, Master Match, Source Properties, Limiting Sensitivity and Convex Hulls.
  - Defines the "CSC Elements" (stacks and ensembles) for the five CSC Modules for processing.
  - Records the status of processing of the CSC Elements in each CSC Module.
  - Regulates the Catalog Automatic Processing (Catalog AP) of the CSC Elements between and within the CSC Modules.
- The HPCC supports the data intensive pipeline processing.

Catalog Automatic Processing

- The original instance of the Automatic Processing system, called "Standard AP", is used to process observation data from the Chandra Observatory. Standard AP:
  - Determines the order of pipeline processing.
  - Prepares the pipeline inputs.
  - Executes the pipeline.
  - Handles the data product outputs.
  - Handles the retrieval and ingest of data products to the data archive.
- "Catalog AP" is an extension of Standard AP. In addition to basic pipeline functions, for CSC 2.0 processing, Catalog AP:
  - Distributes data and pipeline processing on an HPCC.
  - Provides Quality Assurance (QA) intervention and associated re-processing prior to archive ingestion.
- There is an instance of Catalog AP for each of the CSC Modules.

Pipeline Distribution

- A "CSC Stage" is a CSC Module processing component that kicks off one or more pipelines (e.g., Source Properties Master).
- The "Process Manager" distributes the set of pipelines for the CSC stage over the HPCC (e.g., one pipeline per source).
- The HPCC executes these pipelines in parallel.
- For example, all Source Properties Master pipelines for an ensemble are executed in parallel on the HPCC.