Provenance, Planning and Production

Basic Facts

Data come in Time Series of Files

Data produced in discrete time series of jobs

Definitions

Files belong to Data Products:
- Same Data Time Interval in each file, dt
- Same Parameters in each file

Data Sets are Subsets of Data Products:
- Data Sources determine start and end times
- Number of possible files in Data Set = (End - Start)/dt

Data Set Versions are Subsets of Data Sets:
- Error Structure of data in a Data Set Version should be homogeneous

A Climate Data Record is a Single Data Set Version

Derivation

Jobs also come in discrete time series

There are 4 relationships between files and jobs:
1 - File is produced by a job
2 - File is used as input by a job
3 - Job uses a file as input
4 - Job produces a file as output

Files and Jobs form a graph (or network):
1 - Files and Jobs form Vertices (nodes)
2 - Relationships produce Edges (arcs)

Production graph readily formed from relationships
- Keep list of files
- Keep list of jobs
- Keep list of relationships

Common operations correspond to graph traversals

Implications

Error Source Changes:
1. Input Data
2. Source Code
3. Coefficients
4. Connectivity

Lead to → New Data Set Versions

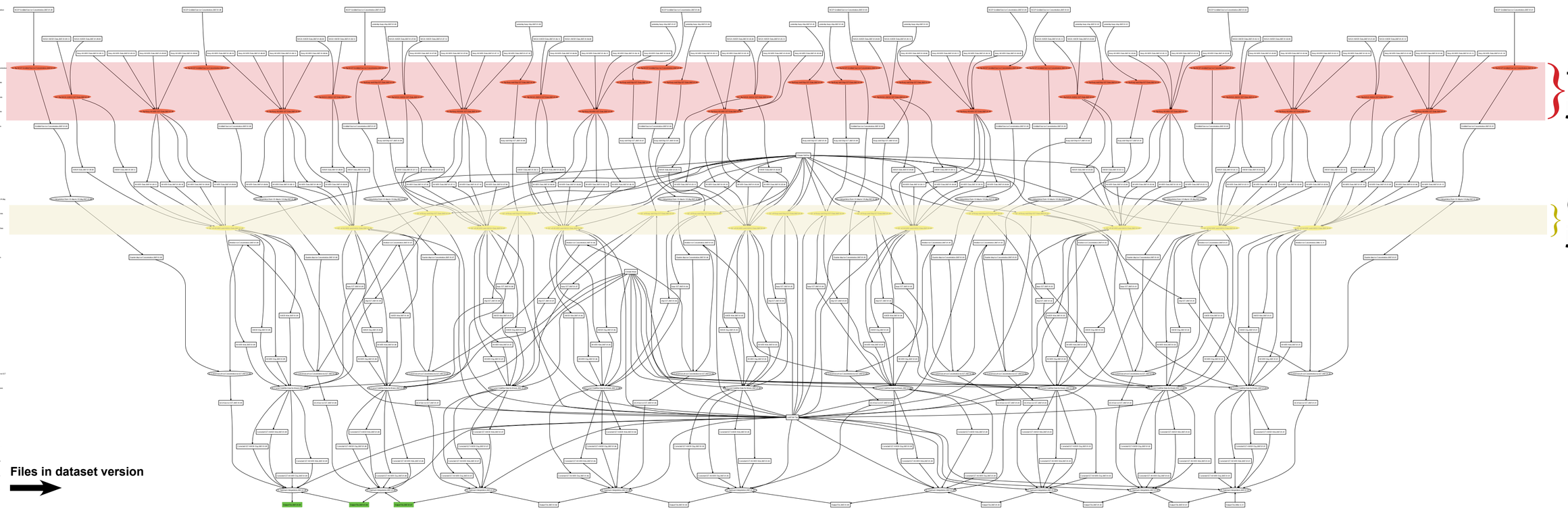
Which are → New Climate Data Records

Create → Reprocessing Cascades

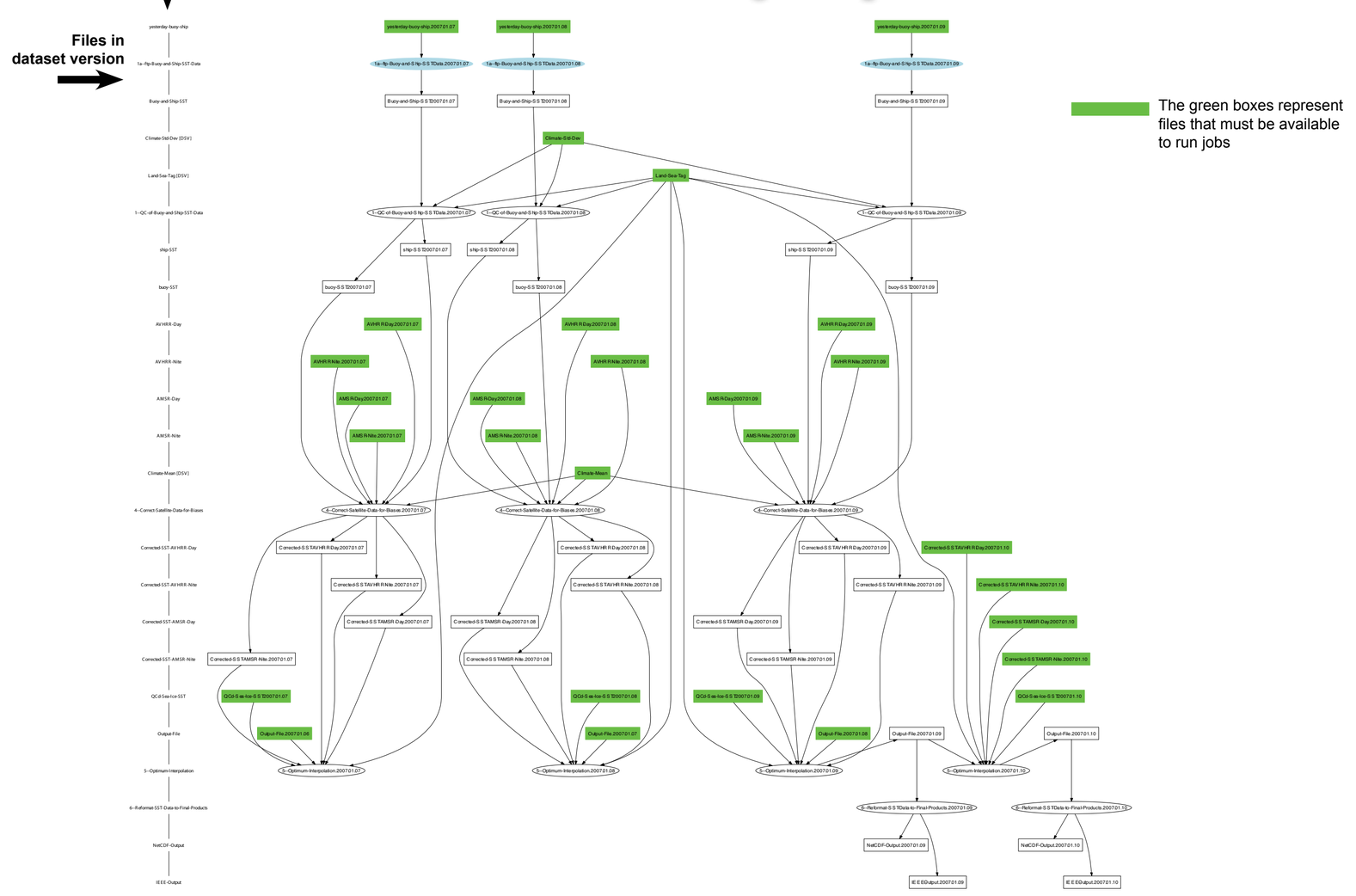
Derived Collection Organization:
Archive
Data Producer
Data Product
Data Set
Data Set Version (Climate Data Record)
File

Allows → Premanent Registration

Provenance Tracking Graph



Production Planning Graph



Bruce R. Barkstrom

NOAA's National Climatic Data Center, Asheville, NC