



Lecture 3

With every passing hour our solar system comes forty-three thousand miles closer to globular cluster 13 in the constellation Hercules, and still there are some misfits who continue to insist that there is no such thing as progress.

- Ransom K. Ferm

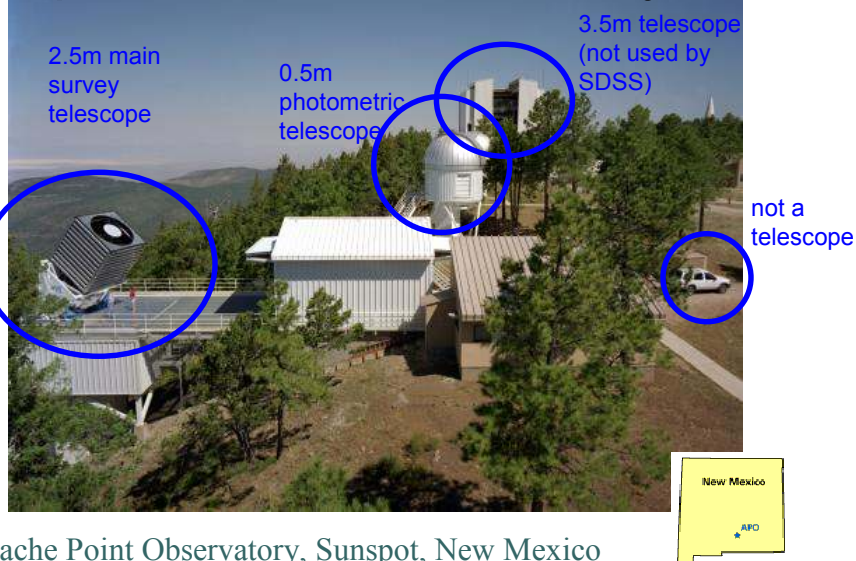


Agenda

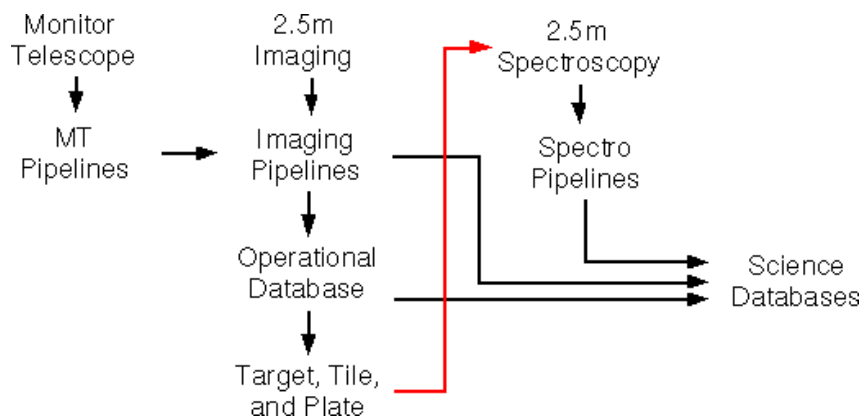
- Homework 1 Questions?
- SDSS Lecture
- Study Questions
- *EOSDIS Demo*

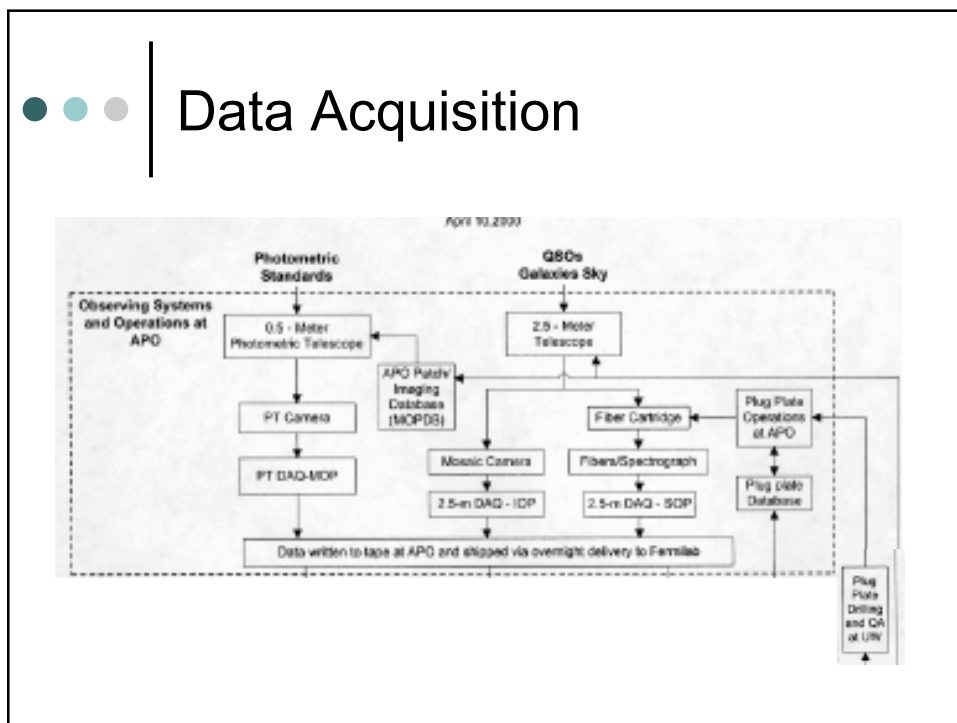
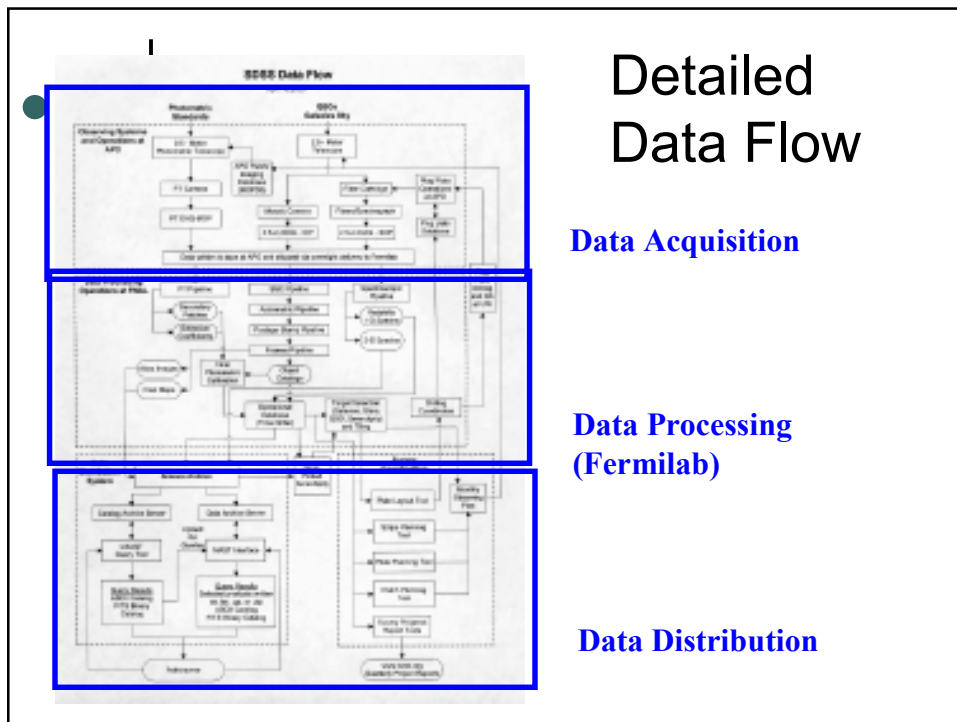


Apache Point Observatory



Coarse Data Flow







Data Acquisition



Good focus area ~
30 full moons



Camera



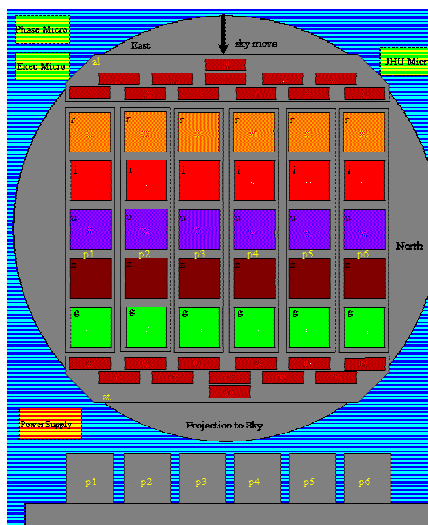
Spectographs



Data Acquisition: 2D Images



- 30 charge-coupled devices (CCDs)
- Each has 4 million pixels
- Each night:
 - 200 gigabytes of data
 - on a dozen tapes

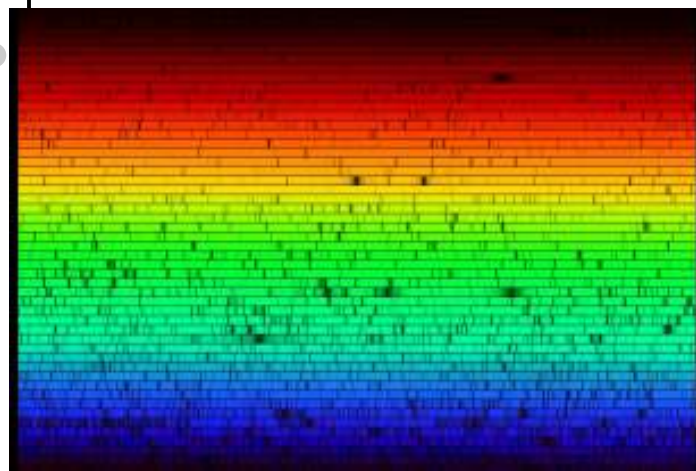
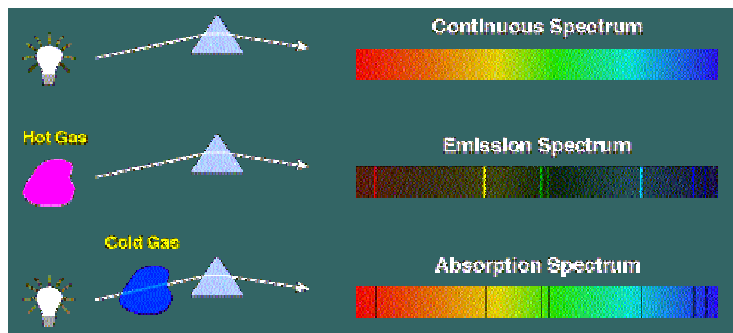




Data Acquisition: Spectra



Data Acquisition: Spectra

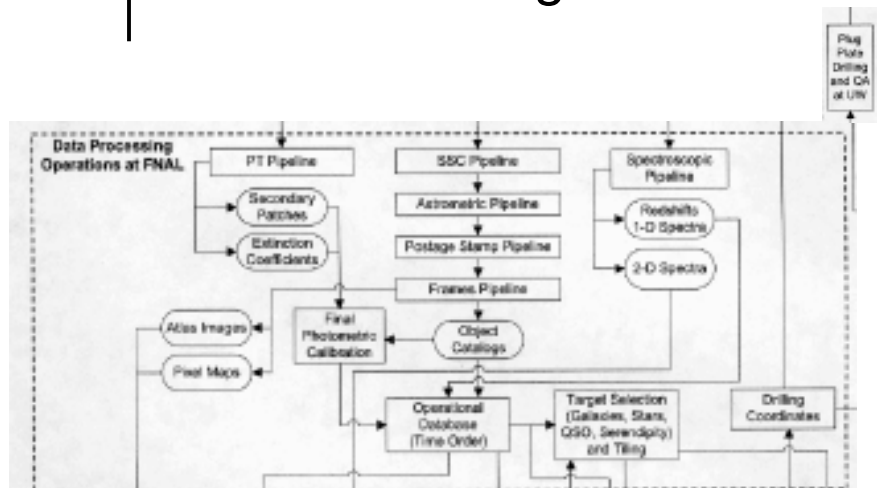


Sun Spectra with absorption lines

Source: National Optical Astronomy Observatory

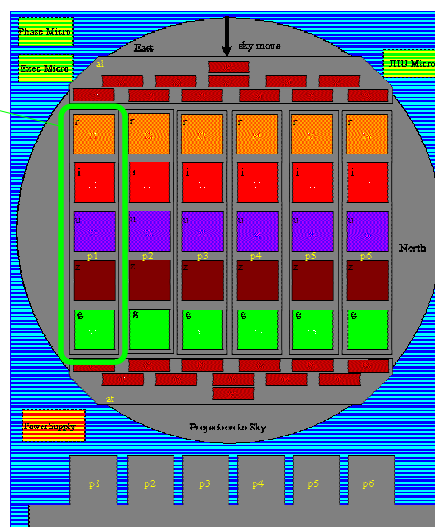


Data Processing



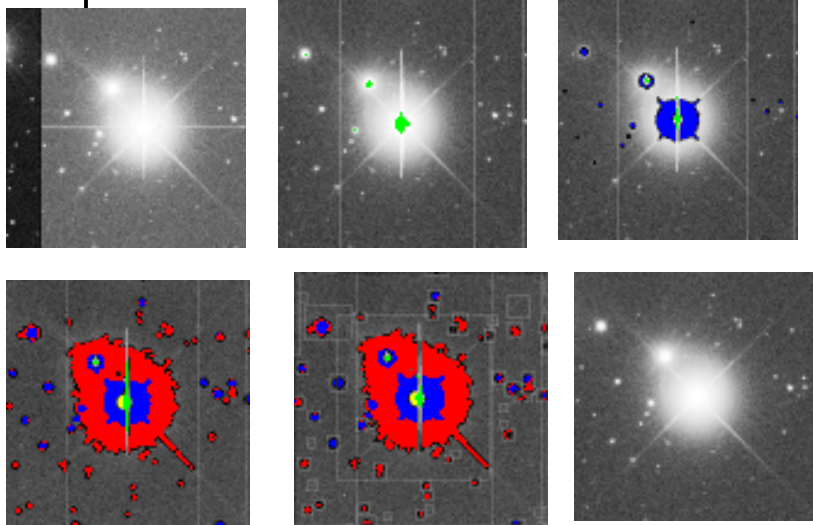
Data Processing

- scanline
- strip = 6 scanlines
- stripe = 2 strips, offset
- frame (per CCD)
 - 2048 x 1489 pixels
 - 10% overlap
- field = frames in all 5 filters





Data Processing: Images



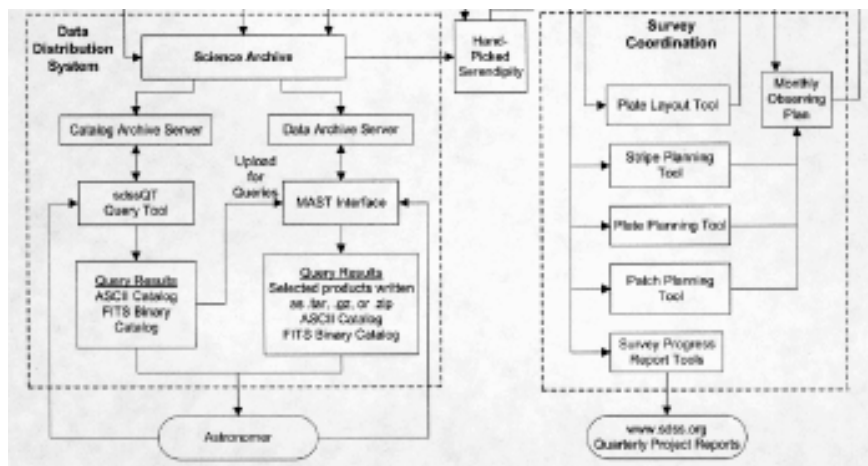
Data Processing: Spectra

- 2D $\rightarrow$ 3D
 - redshift = distance
- Classification
 - Galaxy or Star?
- Wavelengths
 - What substances are involved?

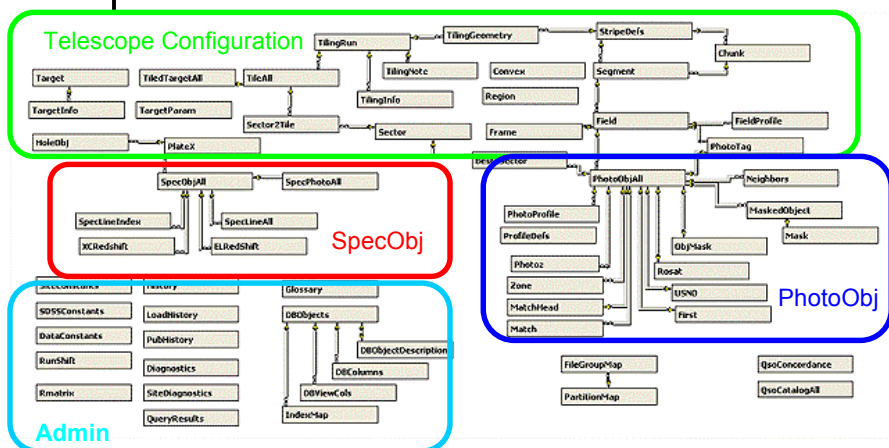




Data Distribution



Data Distribution: Science Database





Data Distribution: Science Database

- 200 million objects (photos, spectra, etc.)
- Numerical attributes in a 100+ dimensional space
- Challenge: how can a relational database scale to large volume of data?



Improving Scalability

- SDSS data too large for one disk or one server
- Base-data objects spatially partitioned across servers
- High-traffic data replicated
- Parallel and distributed query system
- Scan machine – continuously scans dataset and evaluate user defined predicates (partitioned across multiple nodes)
- Hash machine – performs comparisons within data clusters



Overview of SDSS Schema

- SDSS schema browser:
<http://cas.sdss.org/dr4/en/help/browser/browser.asp>
- PhotoObjAll – record describing all attributes of each photometric object
- 100s of columns
- Millions of photos
- Need good indexing/materialized views



SDSS Schema (continued)

- PhotoObjAll table has many views:
 - PhotoObj- all primary and secondary objects
 - PhotoPrimary- all primary photo objects (best)
 - Star
 - Galaxy
 - Sky
 - Unknown
 - PhotoSecondary
 - PhotoFamily (neither primary nor secondary)
- Each view is *Horizontal Partition* (subset of rows)



Other views

- PhotoTag – *Vertical* partition of the PhotoObjAll table (subset of the columns)
- Contains only columns that are most often requested (60 columns, 10% of PhotoObjAll)
- Since rows are smaller (fewer columns), more rows can be loaded into memory and performance improves



Indexes

- Hierarchical Triangular Mesh (HTM)
 - Spatially decomposes region of sky covered by SDSS data
 - Enables faster spatial searches
 - Database indexes
 - Primary key index –primary key of the table
 - Foreign key index -primary key of another table
 - Covering index – index covering one or more columns of a table
 - Speeds up searches if any of the fields included in WHERE clause
- mode, cy, cx, cz, htmID, type, flags, status, ra, dec, u, g, r, i, z, rho*
htmID, cx, cy, cz, type, mode, flags, status, ra, dec, u, g, r, i, z, rho
run, camcol, type, mode, cx, cy, cz



SDSS Database Indexes

- PhotoObj and PhotoTag both indexed
 - 2% subset of PhotoObj
- 50x faster than reading whole PhotoObj table
- 5x faster than reading whole PhotoTag table



Database Size for DR1 (GB)

Filegroups	BESTDR1	TARGDR1
data	1	200
PhotoOther	18.1	
PhotoObjAll	165.4	
PhotoTag	78.1	73.7
PhotoTagIndex	53.6	
PhotoObjIndex	66.3	
PhotoObjProfile	80	
PhotoObjMask	22	17.2
SpecObj	6	
Neighbors	24.2	
Frame	30	30
Log	4.2	2
Total	495.3	322.9

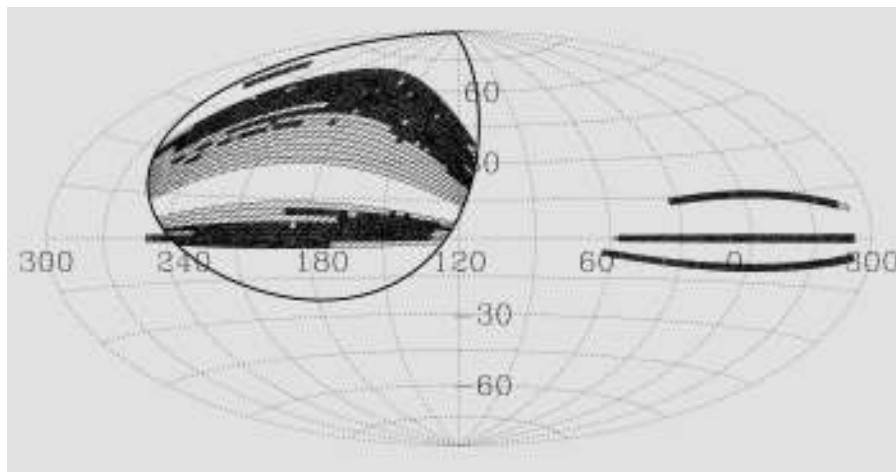


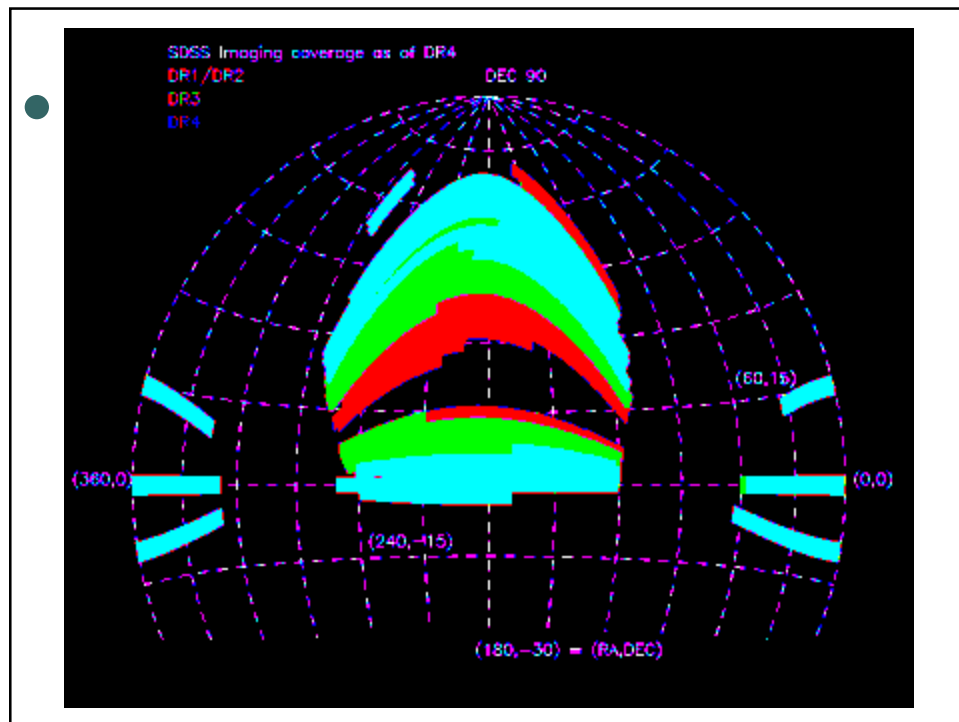
Data Distribution

- CASJobs
 - For long running queries
- Personal Sky Server
 - 1% of total data
 - packaged for one-click install
 - education, testing, demonstrations
- Web services
 - for specific functions



Data Distribution: Releases





Study Questions

