Petasky

http://com.isima.fr/Petasky

Mastodons program of the Interdisciplinary Mission of CNRS

- INS2I

Fusion de **3 projets** Petasky + Amadeus + GAIA

  - ✦ APC (UMR CNRS 7164, Paris)
  - ✦ LAL (UMR CNRS 8607, Paris)
  - ✦ Centre de Calcul de l'IN2P3/CNRS (CC-IN2P3)
- INSU
  - ✦ LAM (UMR CNRS 7326, Marseille)

Petasky: scientific challenges

- Management of scientific data in the fields of **cosmology and astrophysics**
  - ➡ Very large amount of data
  - ➡ Complex data (e.g., images, uncertainty, multi-scales...)
  - ➡ Heterogeneous formats
  - ➡ Various and complex processing (images analysis, reconstruction of trajectories, ad-hoc queries and processing, …)
- Scientific challenges
  - ➡ Scalability
  - ➡ Data integration
  - ➡ Data analysis
  - ➡ Visualisation
- Application context : **LSST project**

From Astronomy to astroinformatics

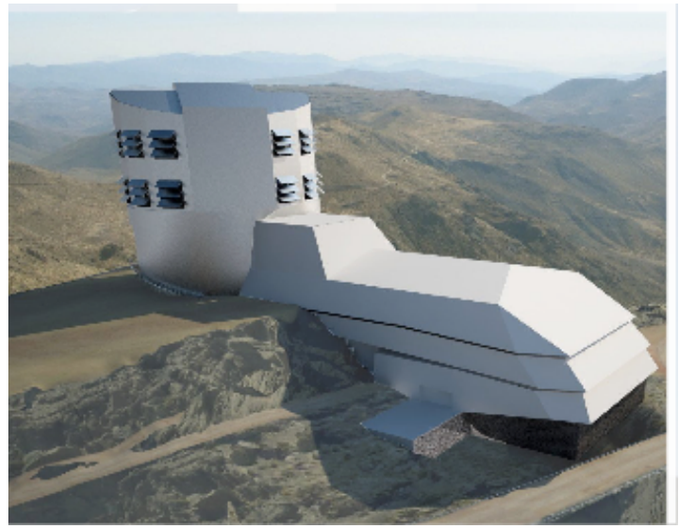
- Modern digital detectors, CCDs,
- Early use of scientific computing, numeric simulations, ..
  - ➡ Antikythera mechanism, between 150 to 100 BC
  - ➡ Supernovae Cosmology Project, 1986
    - 1024x1024 CCD camera, 2 megabytes every five minutes
  - ➡ International Virtual Obse
    - Web of astronomical data
  - ➡ Sloan Digital Sky Survey (
  - ➡ GAIA, launched in 12/201 7/2014
- A culture of sharing data
  - ➡ Data with non-commercial value (more open than healthcare or biomedical science fields)

SDSS

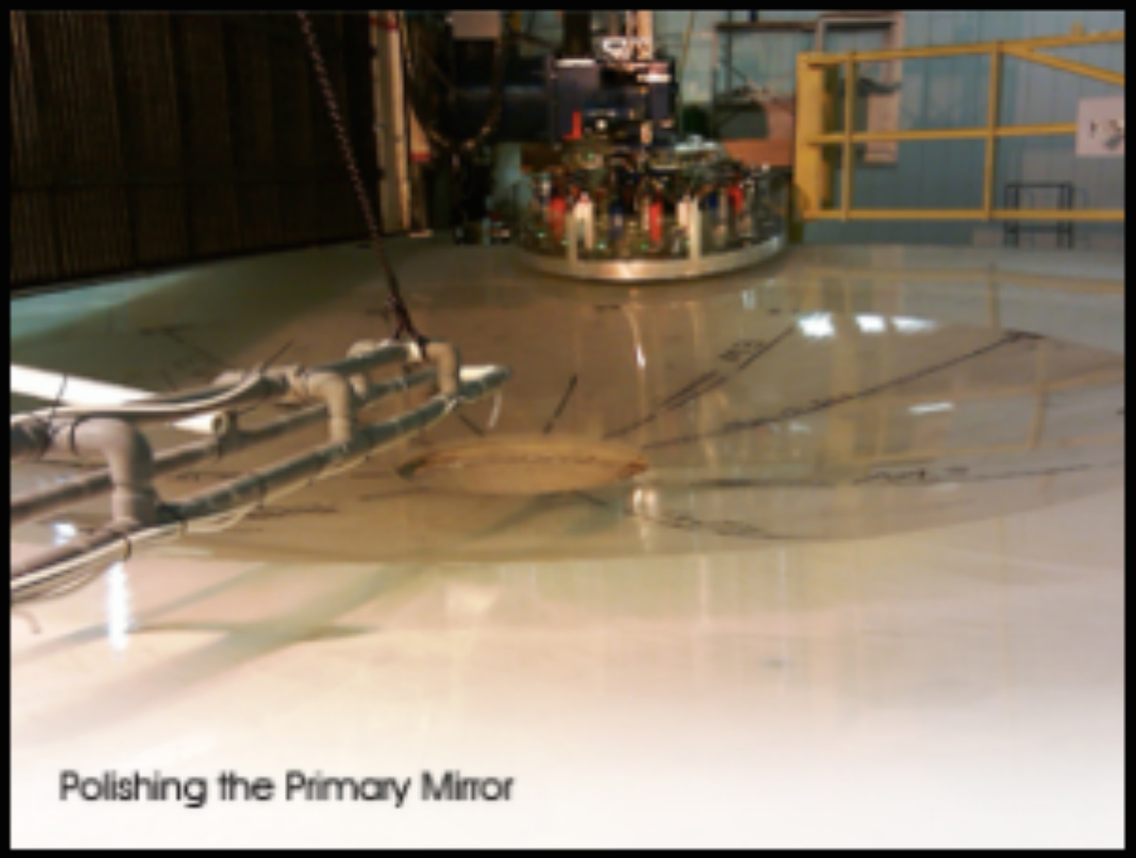
- 2.5 m Telescope, 54 CCD imager
- Operational since 2000
- In 2010, a total archive of **140 TB**

The LSST project
Large Synoptic Survey telescope

A new window over the sky: Telescope of 8.4 m



The New Sky

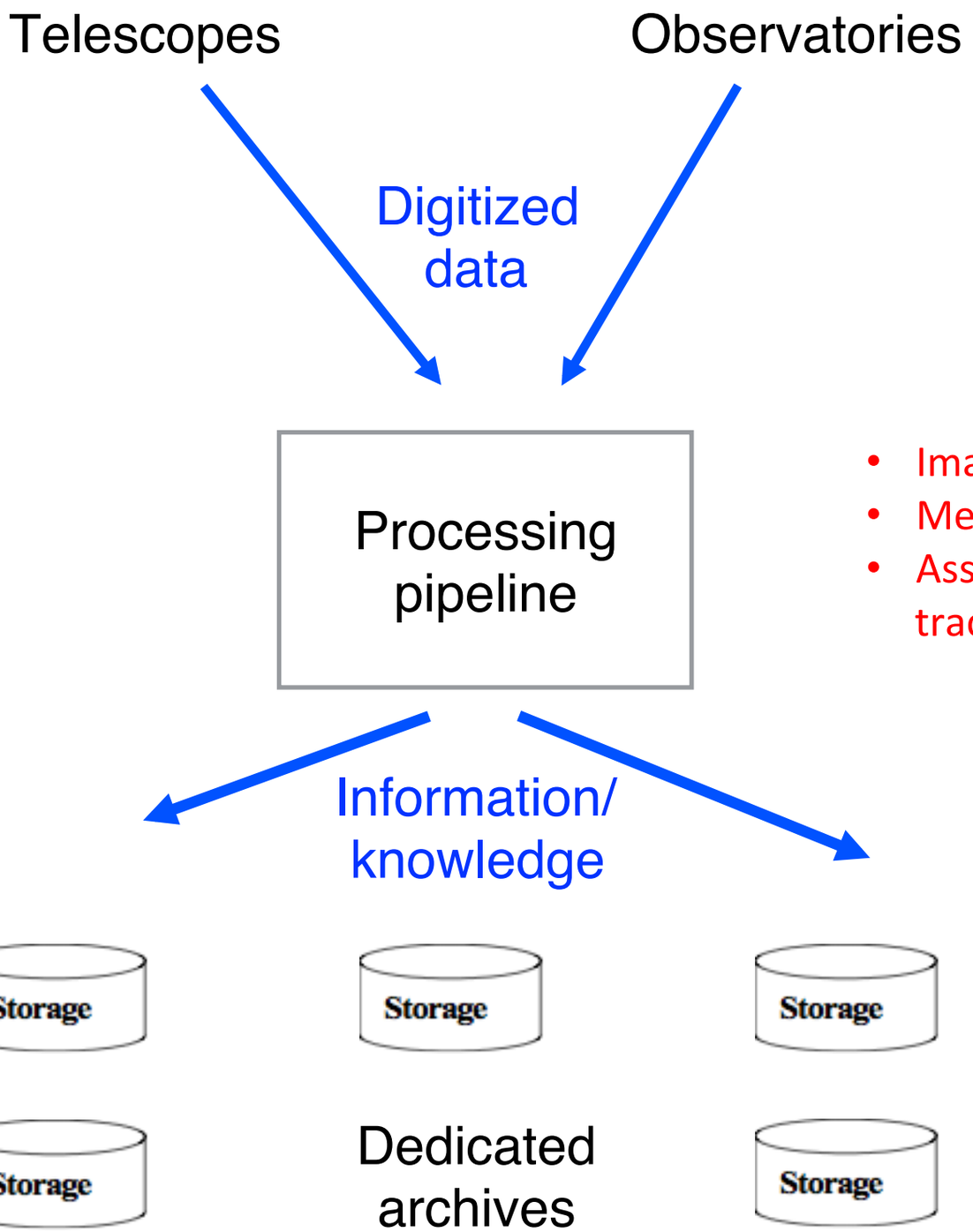


WIDE
A large aperture, wide field survey telescope and 3200 Megapixel camera to image faint astronomical objects across the sky.

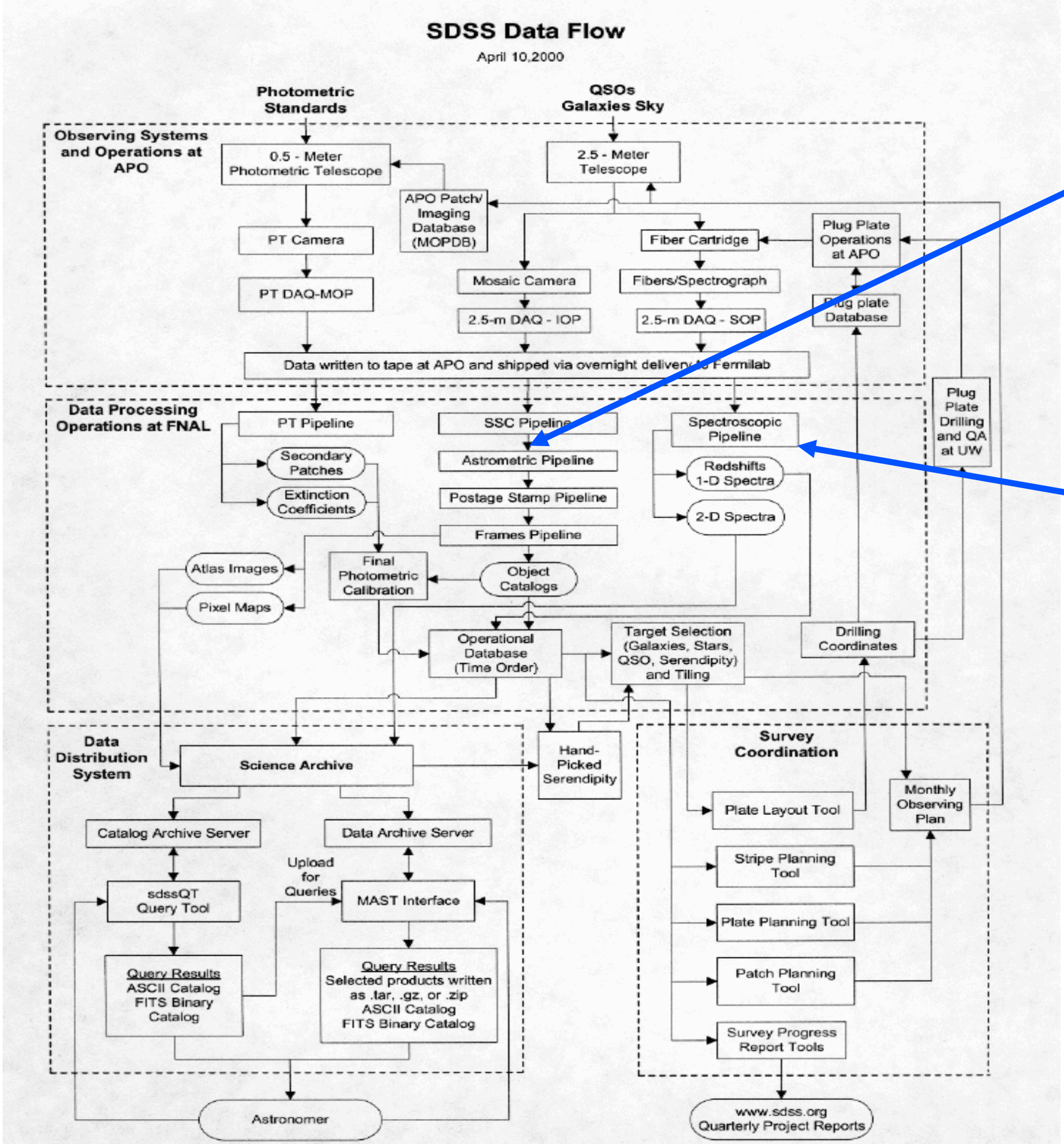
FAST
LSST will rapidly scan the sky, charting objects that change or move: from exploding supernovae to potentially hazardous near-Earth asteroids.

DEEP
LSST's images will trace billions of remote galaxies, providing multiple probes of the mysterious dark matter and dark energy.

Data-driven discovery in Astrophysics



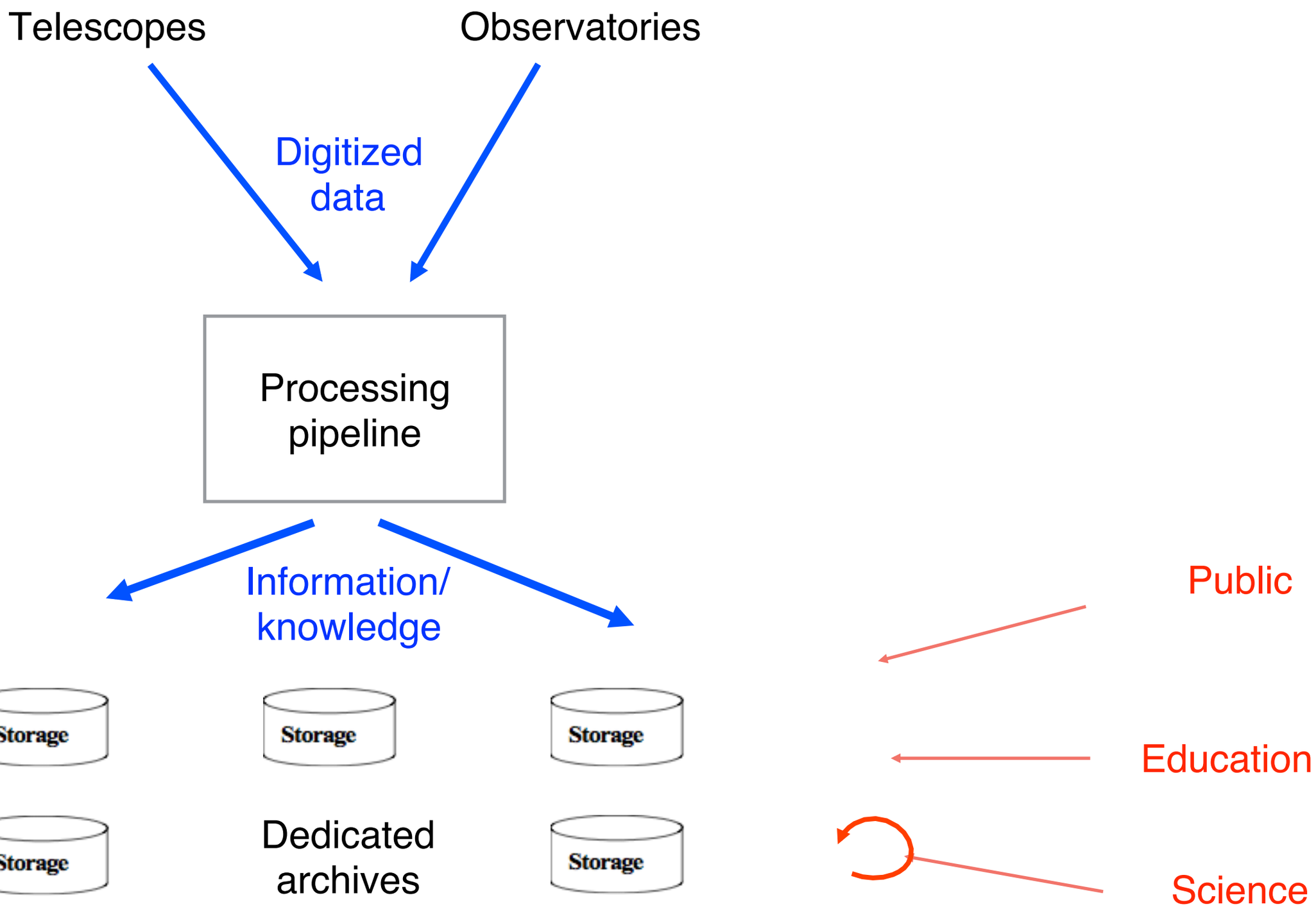
- Image processing
- Measurement of source properties
- Associating sources across space and time, e.g. for tracking solar system objects



Astrometric pipeline
Compute object positions from raw images

Spectroscopic pipeline
redshifts, classification of objects as galaxies, stars, quasars

Data-driven discovery in Astrophysics



Data management challenges in LSST

*"How much the (LSST) project will tell us about our solar system, the dark energy problem and more, will **depend** on **how well we can process the information** the telescope and its camera send back to us - an estimated sum of around ten petabytes of data per year."*

(Mari Silbey, *Space: the big data frontier*, http://www.smartplanet.com/blog/thinking-tech/space-the-big-data-frontier/12180)

*"Plans for sharing the data from LSST with the public are **as ambitious as the telescope itself**"*

> Anyone with a computer will be able to fly through the Universe, zooming past objects a hundred million times fainter than can be observed with the unaided eye. The LSST project will provide analysis tools to enable both students and the public to participate in the process of scientific discovery.

Jim Gray, Turing award 1998

FTP-GREP Model

Remote Archive

FTP

FILES

Analysis

Evolution of data volumes

1986 — Supernovae Cosmology Project: **6,7 KB/s**

2010 — SDSS: **4,3 MB/s**

DR12: **116 TB**

2020 — LSST: **400 MB/s**

Raw data : **60 PB**
Catalog database: **15 PB**

The LSST scientific database

Table	Size	#tuples	#attributes
Object	109 TB	37 B	470
Moving Object	5 GB	6 M	100
Source	3,6 PB	350 B	125
Forced Source	1,1 PB	32 T	7
Difference Image Source	71 TB	200 B	65
CCD exposure	0,6 TB	17 B	45

Data challenges

- Data Volume
  - Typical sky survey may detect 10^8-10^9 sources (stars
  - Multiscale : Space, Time, Wavelengths
- Data quality
  - Missing data
  - …

Table scan : ≈**3h** to scan **1 TB**

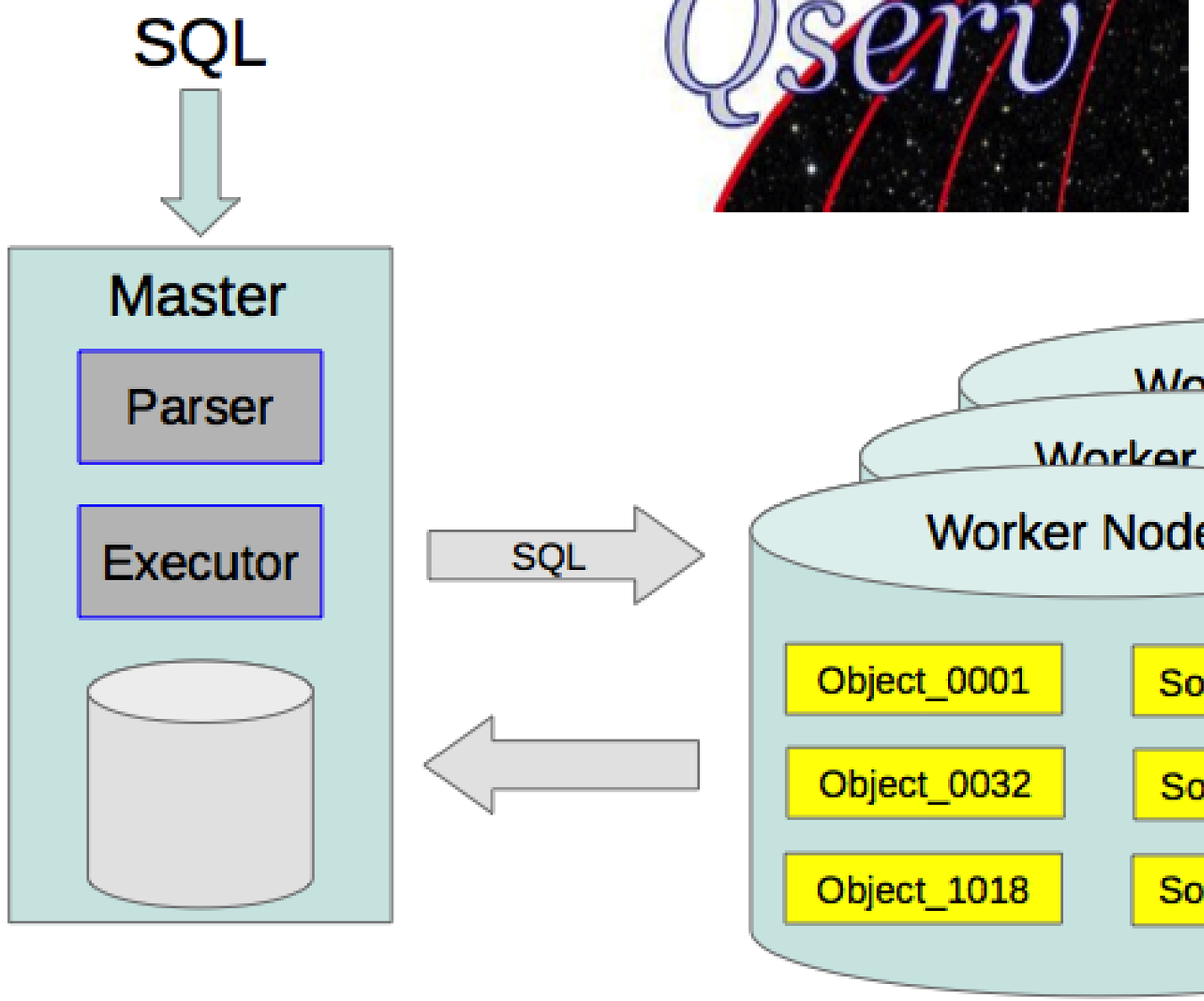
Parallelization
- **< 2 minutes** with 100 HD
- 1TB/sec : 10 000 HD (Google Dremel)
- **1 TB** in less than **1 minute** with Oracle DBMS : 2 RAC nodes + Multiples infiniBand adapters (Maklee)

Modern data management technologies

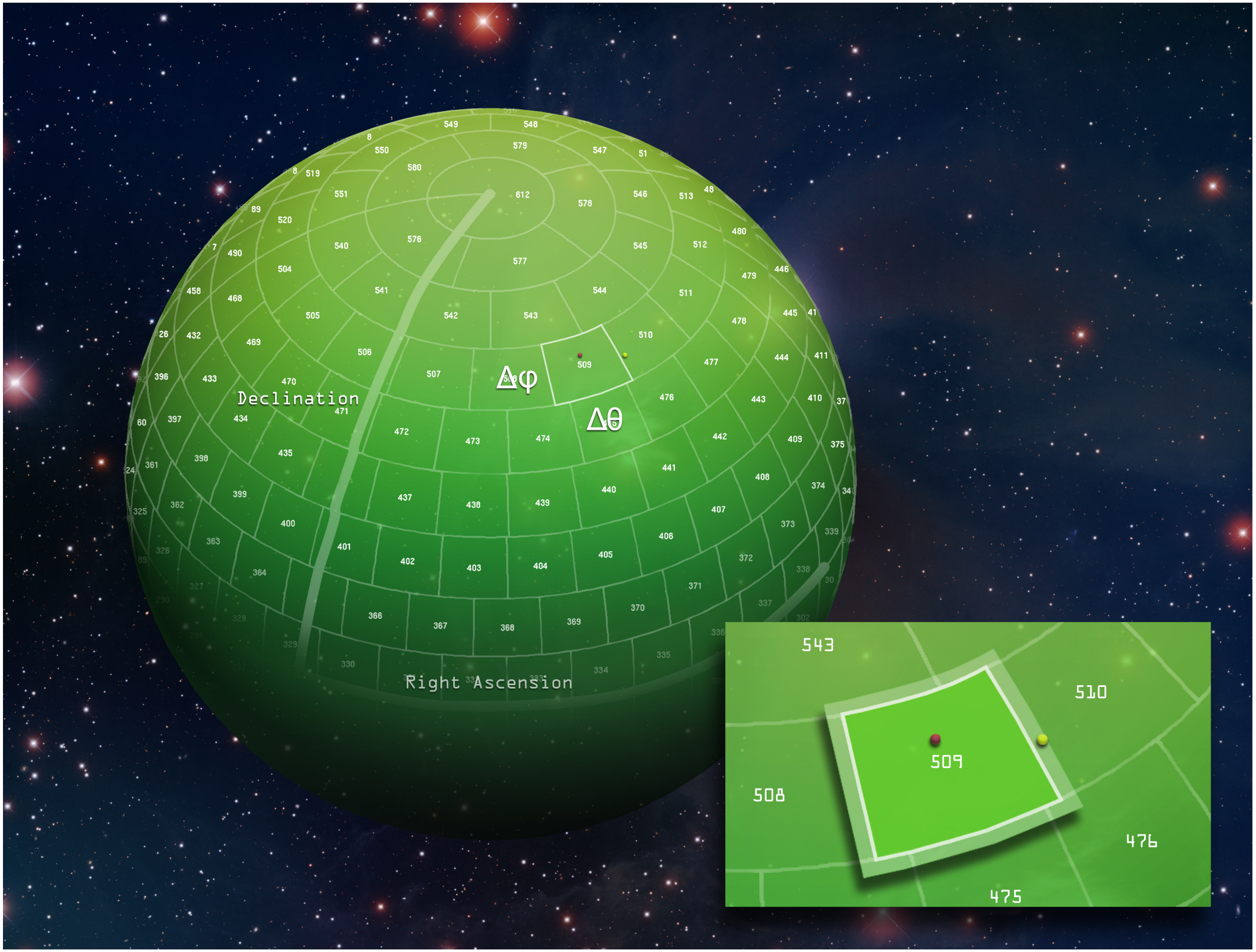
- Massive parallelization
- Virtualization and cloud computing
- Data distribution
- Data storage
  - Row store vs. column store
  - (sophisticated) Indexes
- New computing paradigms
  - Failures resilience
  - Coordination
- Complexity theory and cost models

QSERV Data Management System

- Data partitioning
  - ✓ Horizontal fragmentation of the table Object: $\sigma_{\theta i}(Object)$
  - ✓ Derived horizontal fragmentation of the table Source: $\pi(Source \bowtie \sigma_{\theta i}(Object))$
- Needs for query orchestration



Spatial partitioning



LSST queries per level of difficulty

Supported queries

- Retrieve any type of information about a single object (identified by a given objectId), including full time series.

  SELECT * FROM Object JOIN Source USING (objectId) WHERE objectId = 293848594;

  Few seconds

- Retrieve any type of information about a group of objects in a small area of sky, including neighborhood-type queries.

  SELECT * FROM Object WHERE qserv_areaSpec_circle(1.0, 35.0, 5.0/60)

  ≃1 hour

- Analysing light curves across large area.

  SELECT O.objectId, myFunction(S.taiMidPoint, S.psfFlux) FROM Object AS O JOIN Source AS S USING (objectId) WHERE O.varProb > 0.75 GROUP BY O.objectId;

  ≃1 day (24h)

- Analysing light curves of faint objects across large area.

  SELECT O.objectId, myFunction(V.taiMidPoint, FS.flux) FROM Object AS O JOIN ForcedSource AS FS ON (O.objectId = FS.objectId) JOIN Visit AS V ON (FS.visitId = V.visitId);

  ≃1 week

LSST queries per level of difficulty

Expensive/impossible queries

- **Expensive queries**
  - Find objects far away from other objects (for a large number of objects).

    Question: what is the largest distance we should plan to support for distance based queries involving (a) small number of objects, (b) all objects on the sky?
  - Sliding window queries: Find all 5 arcmin x 5 arcmin regions with an object density higher than rho

- **Impossible queries**
  - Large size results
    - Select all pairs of stars within 1 arc min of each other in the Milky Way region.

Note that queries of these types that operate on small subsets of data wil be supported; ..

    - Join of Source with ForcedSource

Examples of LSST User Defined Functions

- **q3c_ang2ipix(ra, dec) --** returns the ipix value at ra and dec
- **q3c_dist(ra1, dec1, ra2, dec2) --** returns the distance in degrees between (ra1,dec1) and (ra2,dec2)
- **q3c_join(ra1, dec1, ra2, dec2, radius) --** returns true if (ra1, dec1) is within radius spherical distance of (ra2, dec2). It should be used when the index on q3c_ang2ipix(ra2,dec2) is created.
- **q3c_ellipse_join(ra1, dec1, ra2, dec2, major, ratio, pa) --** like q3c_join, except (ra1, dec1) have to be within an ellipse with major axis major, the axis ratio ratio and the position angle pa (from north through east)
- **q3c_radial_query(ra, dec, center_ra, center_dec, radius) --** returns true if ra, dec is within radius degrees of center_ra, center_dec. This is the main function for cone searches. function should be used if when the index on q3c_ang2ipix(ra,dec) is created)
- **q3c_ellipse_query(ra, dec, center_ra, center_dec, maj_ax, axis_ratio, PA) --** returns true if ra, dec is within the ellipse from center_ra, center_dec. The ellipse is specified by major axis, axis ratio and positional angle. function should be used if when the index on q3c_ang2ipix(ra,dec) is created)
- **q3c_poly_query(ra, dec, poly) --** returns true if ra, dec is within the postgresql polygon poly.
- **q3c_ipix2ang(ipix) --** returns a 2-array of (ra,dec) corresponding to ipix.
- **q3c_pixarea(ipix, bits) --** returns the area corresponding to ipix at level bits (1 is smallest, 30 is the cube face) in steradians.
- **q3c_ipixcenter(ra, dec, bits) --** the function returning the ipix value of the pixel center of certain depth covering the specified (ra,dec)

Theoretic vs. paragmatic query complexity

- Tractability revisited
- Coordination
- Scale Independence

  Using Small Data to answer queries on bigdata

Petasky: explored approaches

- Big data management approaches
  - ✓ Distributed and parallel systems
    - ➡ MapReduce-like approaches (shared nothing architecture)
    - ➡ Parallel DBMS (shared all thing architecture)
    - ➡ Spatial partitioning (QSERV for LSST)
  - ✓ Column store DBMSs (Vertica, MonetDB, ...)
  - ✓ Data integration to the rescue
    - ➡ Declarative approach
  - ✓ BSP (Bulk Synchronous Parallel)
  - ✓ Optimisation of spatial queries
- Data mining and knowledge discovery
  - ✓ Interactive exploration of large datasets
  - ✓ Parallel mining of dependencies
  - ✓ New clustering algorithms (One-pass based algorithm, Incomplete enumeration)
  - ✓ Using neural networks to estimate redshift distributions

Experimentation environment

- LSST Dataset
  - SDSS dataset : 3TB
  - Synthetic data: 250GB replicated up to 100TB

    experiments with 2TB-15TB
- Query workload: 35 LSST queries
- System architecture
  - Up to 300 computers
  - Heterogeneity, concurrency, ..

Evaluated systems

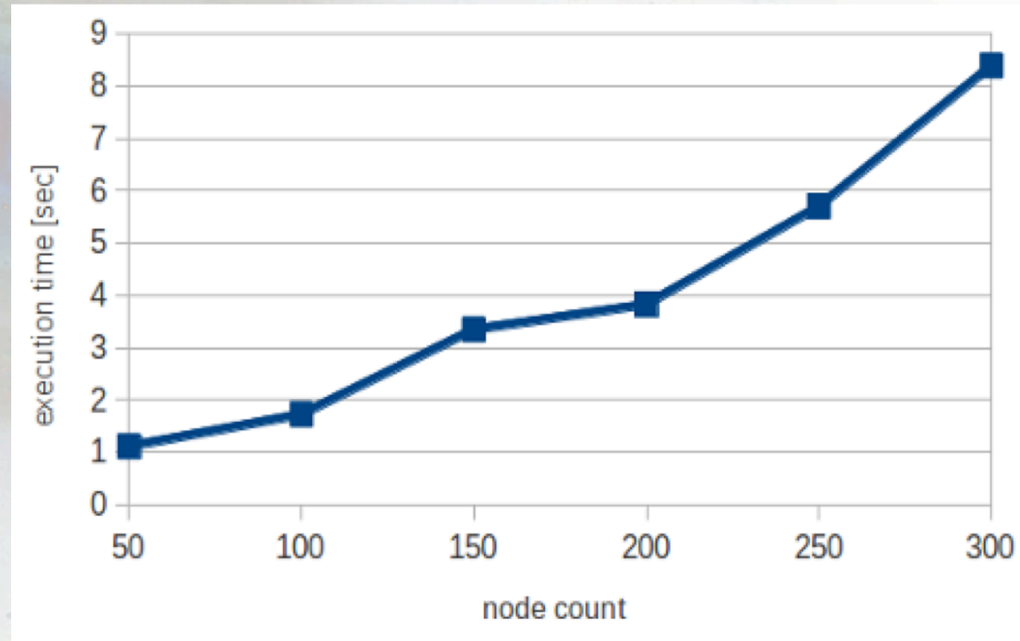
- Qserv
- Hadoop
- HIVE, Hadoop DB
- Relational DBMS : Oracle, Postgres
- MonetDB
- Spatial queries : Spatial Hadoop (MapReduce), PIGEON (PIG), GeoSpark (SPARK)
- Column store and compression methods : SPARK, Pig, Hive et Drill

Experiments with QServ (cont.)

Les études sur Qserv

PetaSky

- **Test de passage à grande échelle** (au CC IN2P3) :
  - 300 nœuds physiques, 120 GB disque, 16 GB RAM
  - 15 Tbytes, 3000 chunks → 50 GB / nœud
    - Parallélisation du déploiement : espace disque = taille DB x2
    - **Test des surcoûts** (tables en cache)
  - Instabilités de performance :
    - 150 → 300 noeuds
    - communications, multi-threading
  - Temps de réponse ~proportionnel au nombre de nœuds



Conclusion sur Qserv

- Requêtes limitées
  - distance ~1 arcmin, non spatial joins
  - Non-partitionable aggregates, User defined functions
- Load balancing
- Ad-hoc query rewriting
- Contraintes liées à un test à grande échelle
  - Nécessité d'une plateforme pérenne
- LPC intégré à l'équipe de développement de Qserv

MapReduce-based approaches: (provisional) conclusion

It is not **THE ultimate solution** for big data

- Data Loading
- Data partitioning
- Suited for Embarrassingly parallel workloads
- Sophisticated optimization techniques (indexing, partitioning, Buffering, ...) are still missed

Learned lessons and research directions

No one fits all

- MapReduce-based algorithms can be useful to implement physical operators
- Hybrid system: row/column store
- Need for more research on
  - ✓ Abstraction adequate to the scientific domain
    - Array data model (SCiDB)?
  - ✓ Support of user defined functions
  - ✓ Optimization techniques embedded in the data management system
  - ✓ Scalability of information integration framework

Petasky: data management challenge

Techniques to build an **efficient** and **easy to use** data access and analysis system at a **reasonable cost**

From	→	To
• Specialized Hardware		• Commodity machines
• Programming		• Querying
• Ad-hoc optimization		• **Generic** system

Need for more research on

- ✓ Abstraction adequate to the scientific domain
  - Array data model (SCiDB)?
- ✓ Support of user defined functions
- ✓ Optimization techniques embedded in the data management system
- ✓ Scalability of information integration framework and datamining techniques

Working cross disciplines

… working cross cultures

Scientists

Data mining/learning/statistics

Foundational approach to build scientific data management system

Data management

Visualisation

Emergence d'une communauté interdisciplinaire ..

- Projet européen COST BigSkyEarth
- Co-encadrements de thèses
- Plateforme Glactica (https://galactica.isima.fr)
  - IR : Programme PlaSciDo, INS2I
  - Equipement : CPER région Auvergne
- Groupe de travail Maestro
  - MAsses de données En aSTROnomie et astrophysique du GDR Madics (http://www.madics.fr/actions/actions-en-cours/maestro/)
- Journées Plateformes (https://indico.in2p3.fr/event/13365/)
  - 6-7 octobre 2016, Clermont-Ferrand

dépasser les frontières

MERCI

A multidisciplinary research program

- Astronomy and astrophysics
- Computer science/applied mathematics
  - Machine learning and data mining
  - Bigdata management
  - Visualization
  - Distributed computing
- Education
- Drive innovations in industry

Space of solutions and associated challenges

! Clearly beyond the capacities of centralized systems

- Distributed and parallel systems
  - ✓ Data distribution
  - ✓ Computation distribution
  - ✓ Failure resilience
- Storage model
  - ✓ row store vs. column store
  - ✓ (sophisticated) Indexes
- Benefit from modern hardware
- Scalable datamining and machine learning techniques
- Complexity theory and cost models
  - ✓ Standards measures: I/O, data transfer, ..
  - ✓ Cost of coordination

What makes DB technology successful in business domain?

- **Abstractions**
  - ✓ *Relation* instead of files, blocks, tablespaces, segments, extents, access path
  - ✓ Relational algebra instead of algorithms
- **Declarative** query language
  - ✓ Express what you want not how to get it
- **Optimization**
  - ✓ Rather naive techniques but enough for the business world