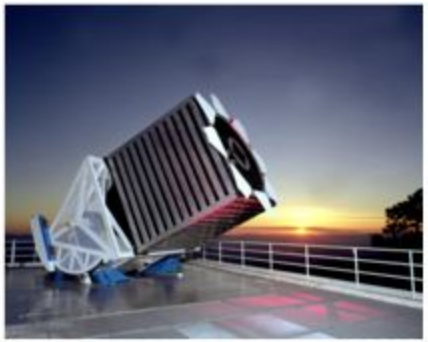
Processing of raw astronomical data of large volume by MapReduce model

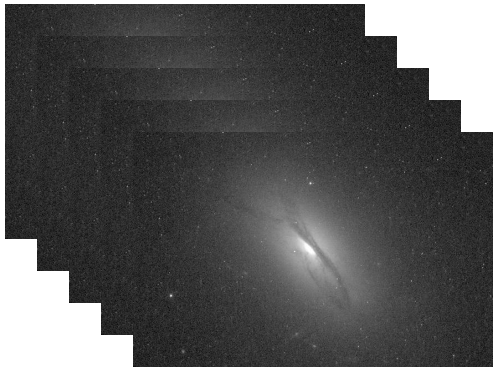
Gerasimov S.V., Kolosov I.Y., Glotov E.S., Popov I.S.
Faculty of Computational Mathematics and Cybernetics Lomonosov Moscow State University

Meshcheryakov A.V.
Space Research Institute of the Russian Academy of Sciences

Digital Sky Surveys



Telescope



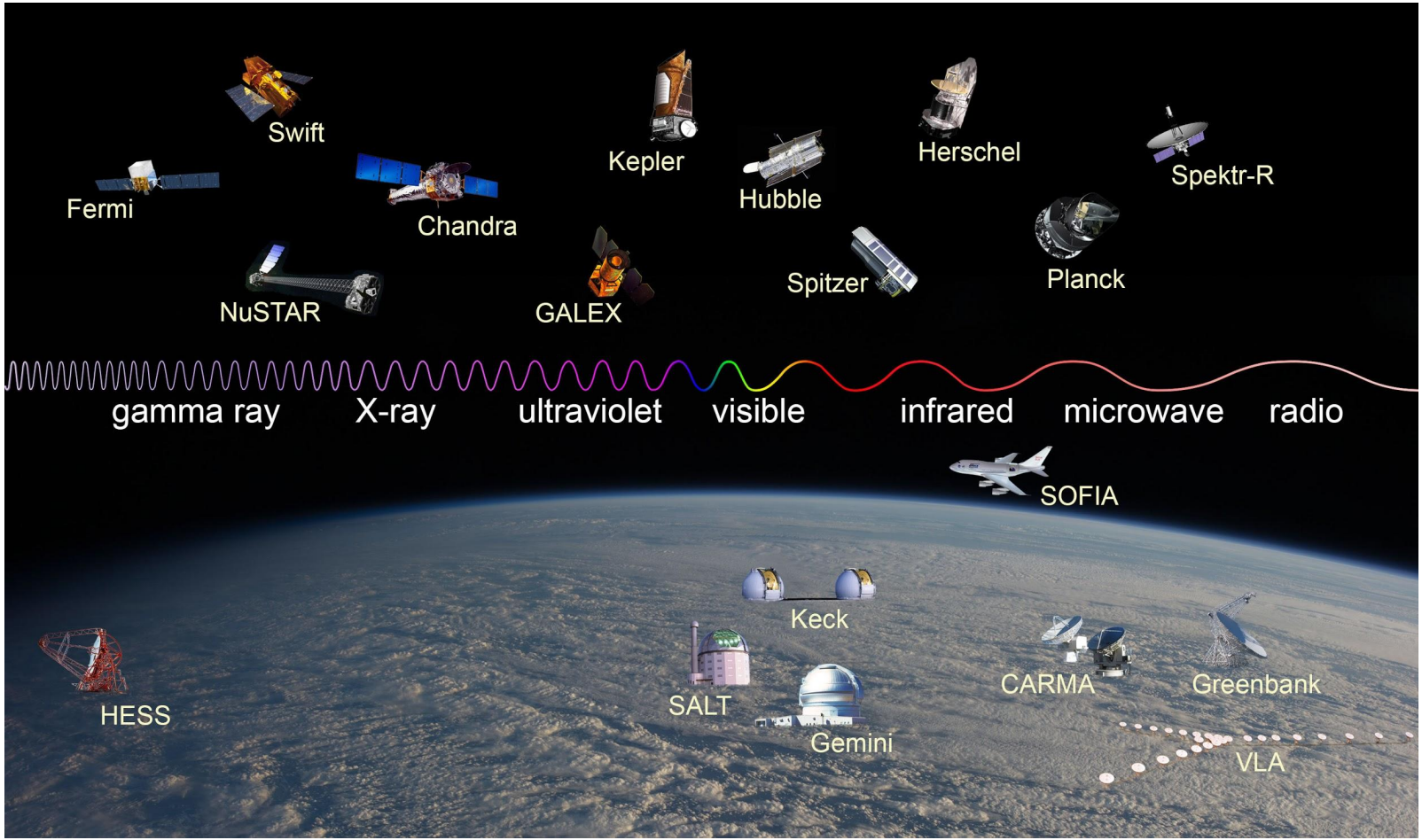
Raw images

pipeline

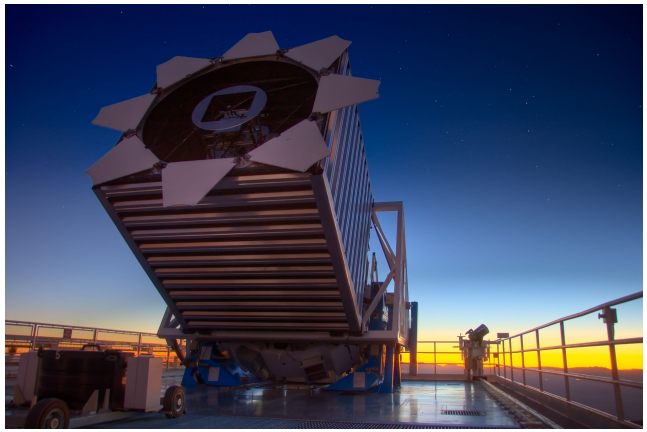
objId	psfmag_u	psfmag_g	psfmag_r	psfmag_i	psfmag_z
1237666660766310000	23.19847	23.39965	22.43144	21.94284	21.60491
1237660237641480000	22.87714	22.1942	21.17471	20.57603	20.15581
1237663083583700000	26.06604	22.77593	22.0157	22.258	21.91062
1237678919666230000	23.07942	22.75924	21.75666	21.05289	20.83766
1237680263998660000	25.48579	24.43595	23.14009	22.16851	21.72804
1237679503780540000	23.47491	22.81789	22.27086	21.90896	21.73548
1237662963319370000	24.63494	23.20006	21.72305	21.26964	20.49038
1237678602377490000	24.88241	24.08308	23.08055	21.8592	21.36703
1237666494327480000	24.39738	23.70336	22.05684	21.53591	22.26463
1237653438159580000	23.42393	21.93365	20.68204	20.25468	19.96516
1237649770787170000	25.59212	19.82089	18.37598	18.10021	24.1325
1237663234446060000	23.74508	23.08823	22.34327	22.33628	21.95397

Catalogue

Multiwavelength astronomy



Data volumes



1998-2009

- 120 mega-pixels
- 0.08PB totally

2010-2014 (PS1)

- 1400 mega-pixels
- 0,4 PB / year
- 2PB totally

Data volumes

THE DARK ENERGY SURVEY

2013-2018

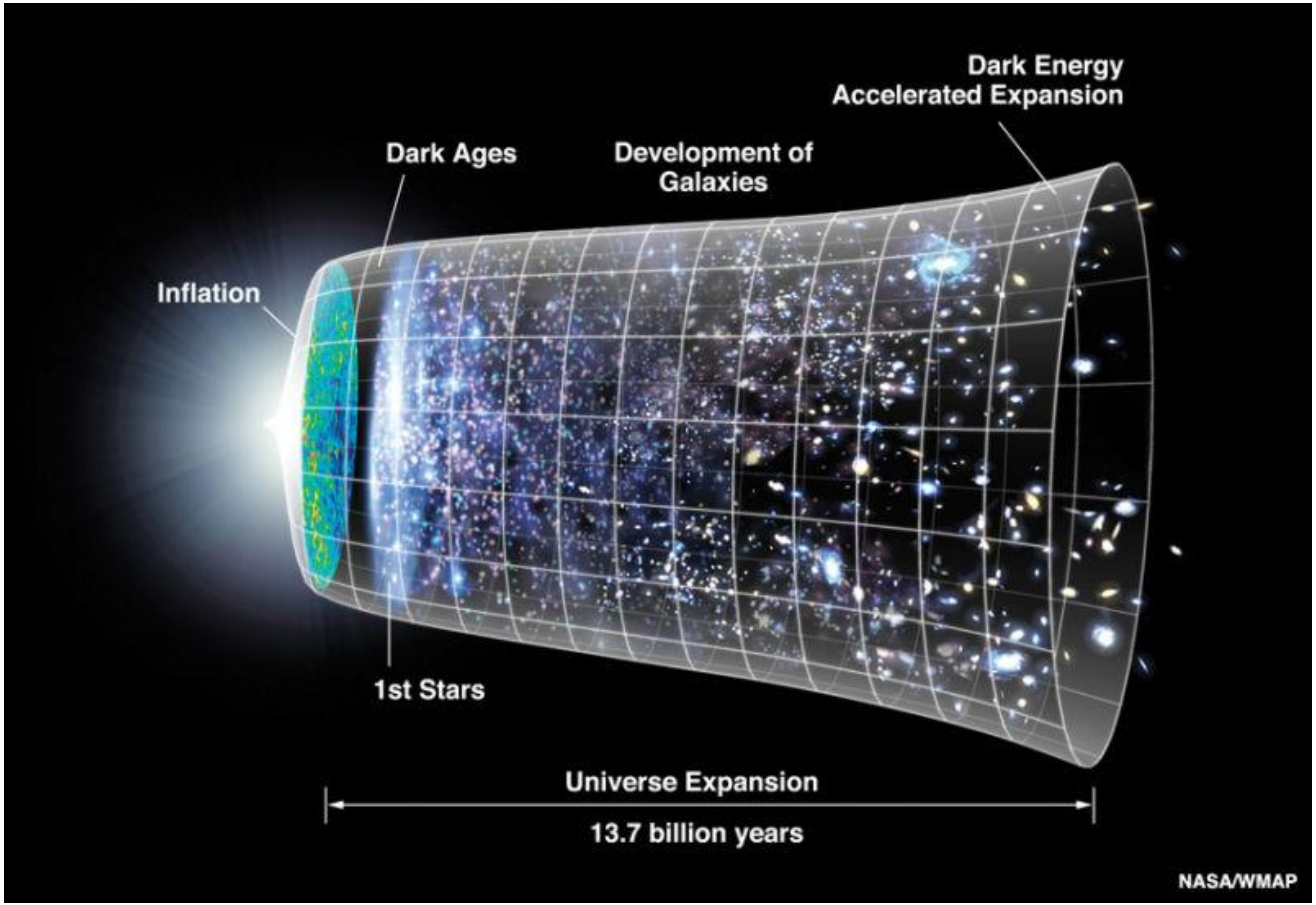
- 570 mega-pixels
- 0.1 PB/year
- 0.5 PB in total

LSST
Large Synoptic Survey Telescope

2022-2032

- 3200 mega-pixels
- 15 TB data every night
- 60PB of raw data in 10 years
- 15PB catalogue

The astronomical science of big data sets



Biggest questions

- nature of dark energy
- nature of dark matter

Small effects

- requires large volumes (all-sky and high depth imaging)
- systematics are important

Small teams

- require ability to analyse big data sets

Our research

- Research of big data technologies to process and store raw astrophysical data (current report)
- Creation of experimental prototype of configurable astronomical image data pipeline based on MapReduce (current report)
- Research & developement of machine learning algorithms and their “big” versions to solve actual astrophysical tasks:
  - extragalactic objects distance (and other properties) estimation
  - star/galaxy/quasar classification
  - transient sky objects detection and classification
  - exploration of hidden structure of sky objects distribution

… to help astrophysicists to do their job better!

Pipeline steps

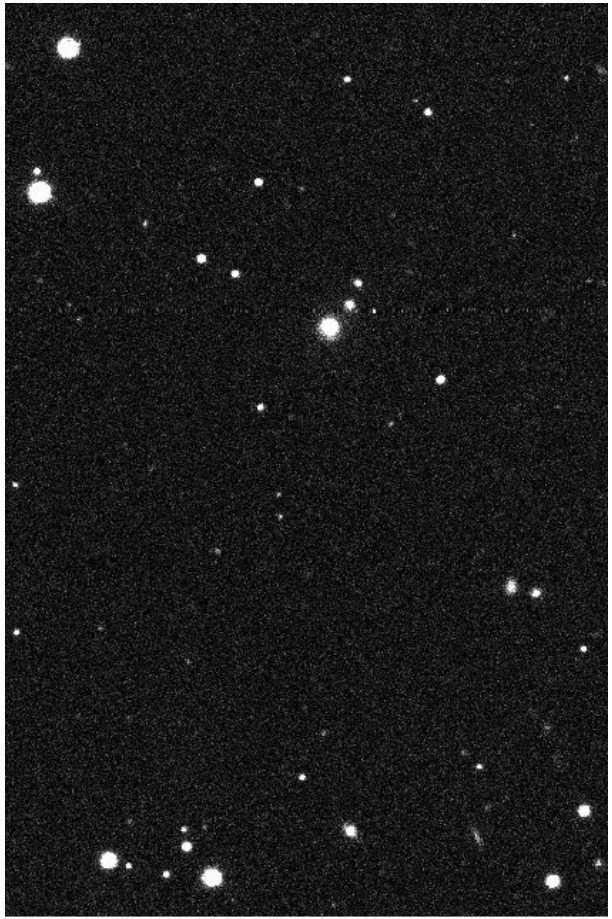
Telescope / sky survey side steps

Step	Description
Raw processing	Removal of CCD noise and artefacts
Astronomical calibration	Detection of World Coordinate System (WCS)
Photometric calibration	Objects intensity callibration

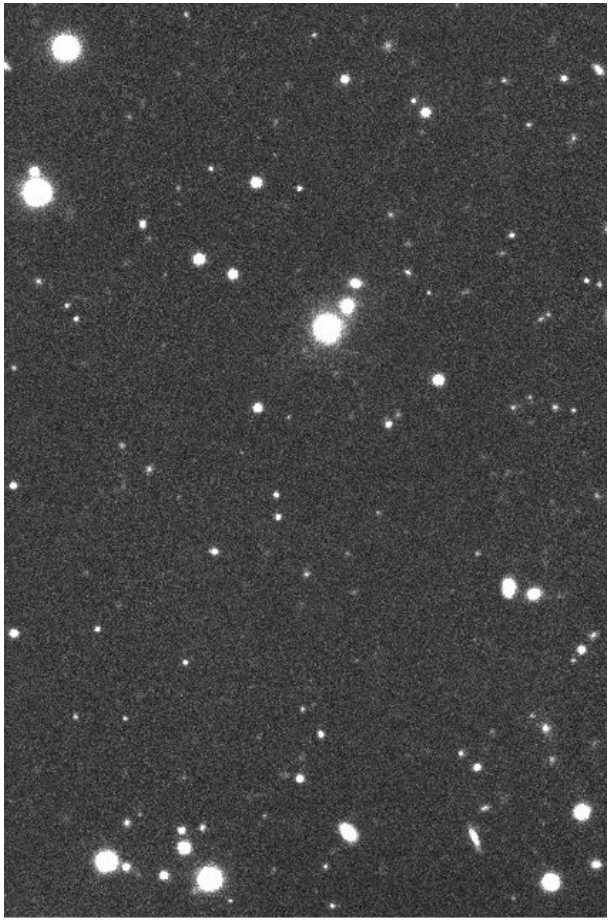
Intelligent steps

Step	Task
Co-addition	Projection
	Background estimation and subtraction
	Stacking
Catalogue creation	Background estimation and subtraction
	Image areas filtering
	Image segmentation (object groups extraction on filtered images)
	Deblending (object extraction inside groups)
	Artifacts removal (cleaning)
	Basic objects features measurement
	Star/galaxy classification
	Creation of PSF-model on stars
	Accurate measurement of all features of objects taking into account PSF-model

Co-addition depth effect



+ 53 more images



Pipiline tools: astromatic.org

SWarp

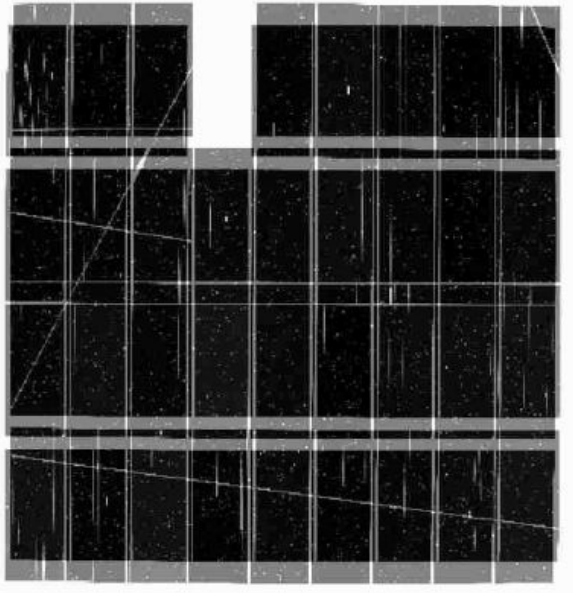
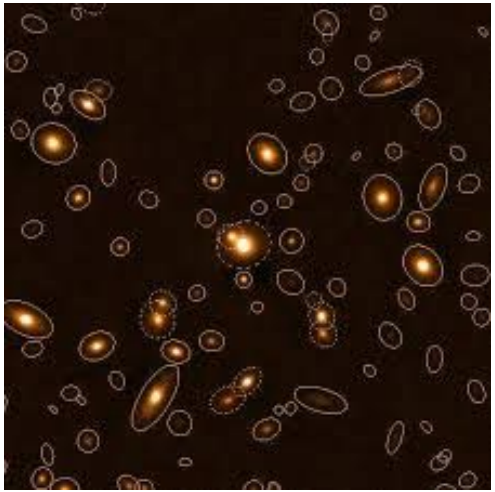


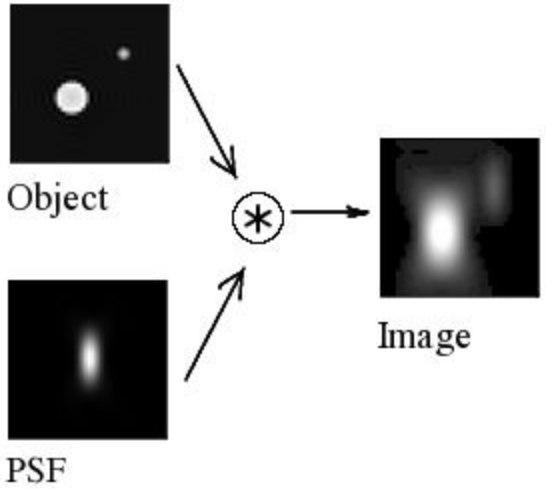
Image co-addition

SExtractor



Objects extraction, measurement of their features

PSFEx



Point Spread Function (PSF) -modelling

Impact of telescope and atmospheric effects on astronomical images

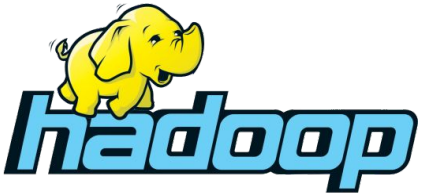
PSFEx uses stars detected on image for empirical PSF modeling.

Courtesy of LSST PhotoSIM

Distributed pipelines

- Astromatic-Wrapper https://github.com/fred3m/astromatic_wrapper
- Wiley et al. Astronomy in the Cloud: Using MapReduce for Image Coaddition
- Farivar et al. Cloud Based Processing of Large Photometric Surveys
- Montage: an astronomical image mosaic engine http://montage.ipac.caltech.edu/
- LSST www.lsst.org http://confluence.lsstcorp.org/

Proposed approach

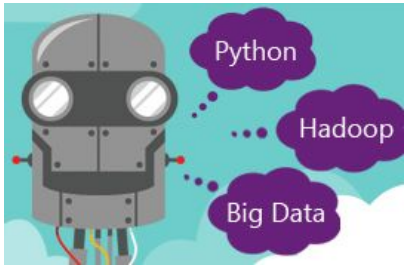


pipeline

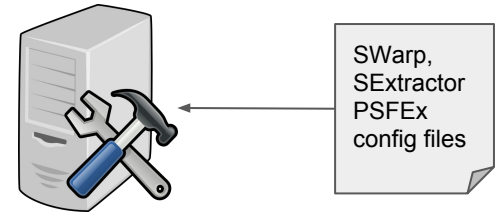
catalogue

machine learning, data mining

Microsoft Azure for Research

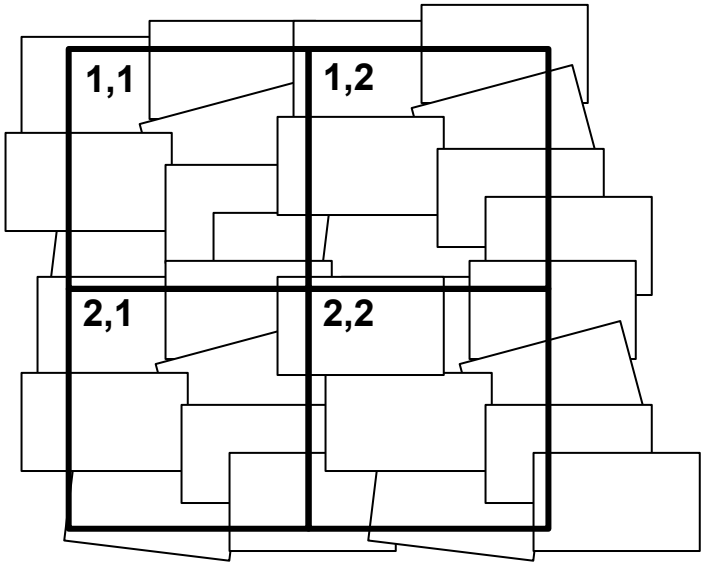


Proposed approach



MapReduce + HDFS

Co-addition step



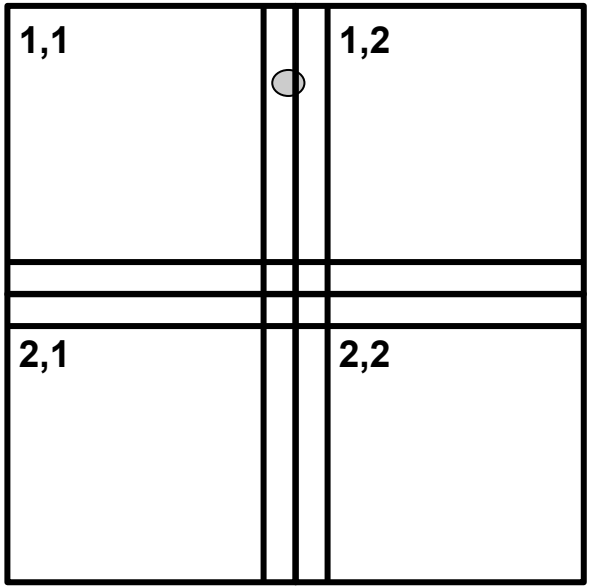
map(filename, image):
  if doesn't belong to target sky fragment
    return
  image'=projection_and_background_subtraction (image) # SWarp
  return ((i,j), image')

reduce((i,j), list<image>):
  cell=stacking(list<image>) # SWarp
  return ((i,j), cell)

Catalogue creation step

map((i,j), stacked_image**):**
  basic_object_features, object_icons=basic_extract(stacked_image) #SExtractor
  psf_model=create_psf_model(basic_object_features, object_icons) #PSFEx
  object_features=extract(stacked_image, psf_model) #SExtractor
  return ((i,j), objects_features)

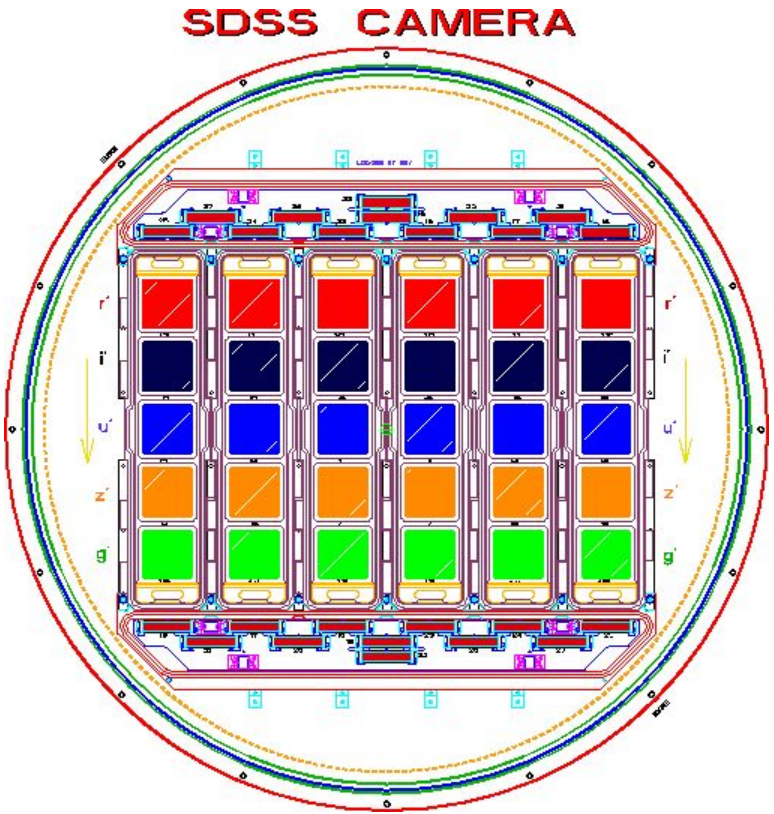
Objects on cell frontier



Experiments

Stripe82 SDSS DR 12

- stripe
- run (north / south)



Experiments

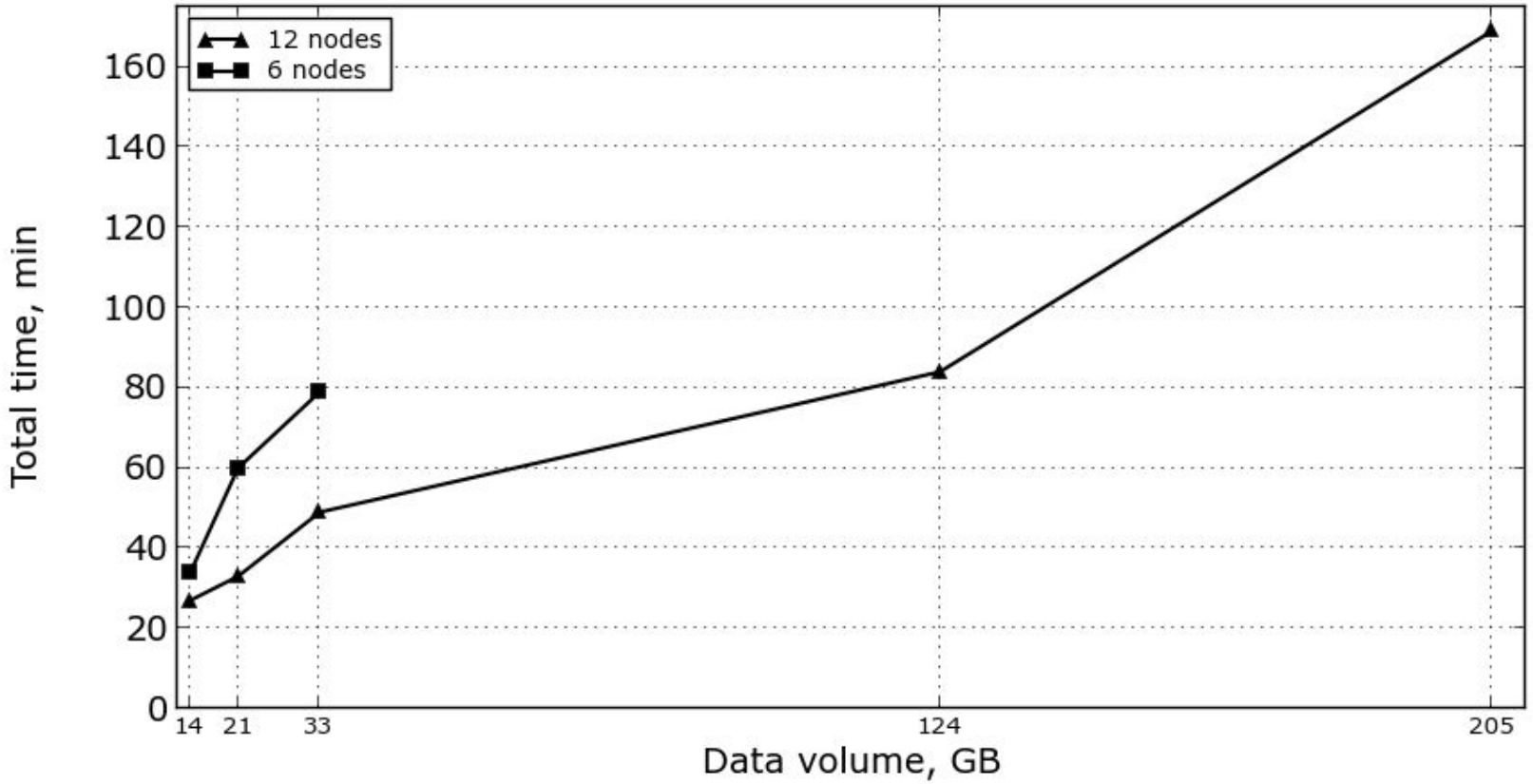
D12 (4 cores, 28GB RAM, 200GB SSD) x 12

Co-addition
cell-size: 0,7°x0,7° (0,5°+0,2°)
cell-count: 60x5
mapreduce.map.memory.mb=5GB
mapreduce.reduce.memory.mb=5GB

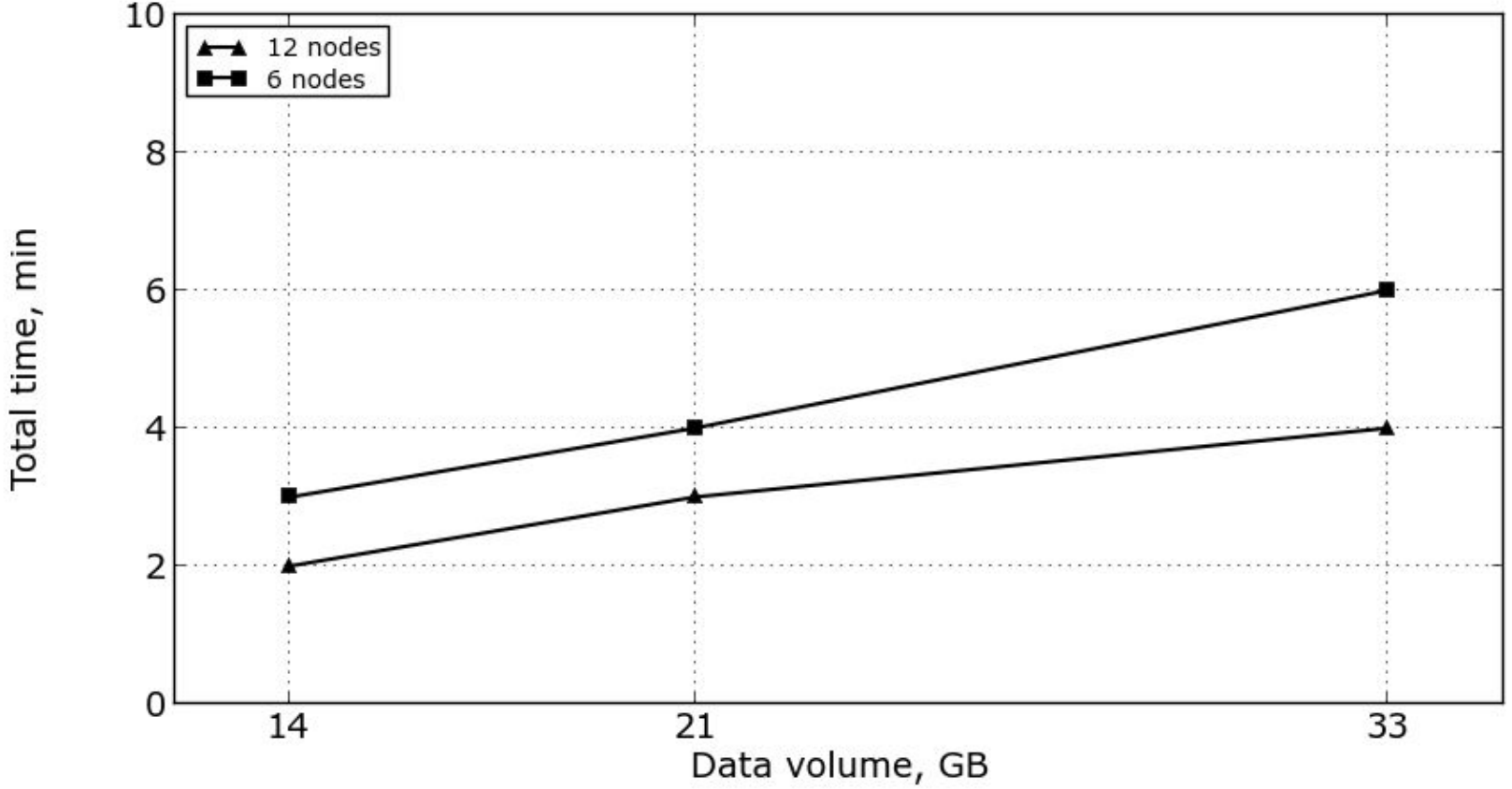
Catalogue creation
mapreduce.map.memory.mb=5GB

Source FITS-files (size of each ~12MB) were initially converted to SequenceFile format to suite HDFS block size.

Scalability: co-addition



Scalability: catalogue creation



Co-addition: MapReduce metrics

Volume	Processing time, min.				
	map	shuffle	sort	reduce	total
14 GB	16	17	1	10	27
21 GB	17	17	1	14	33
33 GB	23	27	1	21	49

Results & next steps

- Current numbers showed MapReduce pipeline is practically useful for astrophysicists
- Further experiments scheduled to measure scalability on larger data volumes (1+TB)
- Some enhancements in the MapReduce pipeline will be implemented and their impact on performance will be researched
- 2016-2017: research and development of “big” versions of machine learning and data mining algorithms for astrophysics

Aknowledgment

The project is supported by RFBR grant #15-29-07085 офи_м

Cloud resources were provided as grant “Microsoft Azure for Research”

Experiments: catalogue creation

Volume	Processing time
15.7GB	2min 32s
23.5GB	3min 24s
34.1GB	5min 45s
45.3GB	7min 5s