SciPIC and CosmoHub

Massive cosmological data generation, analysis and distribution using a Big Data platform

Study case: Flagship mock galaxy catalog

Pau Tallada, Alex Alarcón, Linda Blot, Jorge Carretero, Jordi Casals, Francisco Castander, Marc Caubet, Pablo Fosalba, Santiago Serrano

PIC
port d'informació científica

ICE CSIC-IEEC

IFAE
Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST
Barcelona Institute of Science and Technology

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

Contents

- Euclid @ SDC-ES (hosted by PIC)
- Flagship data regime
- PIC Big Data Platform
  - SciPIC
  - CosmoHub
- Conclusions & Future work

Port d'Informació Científica (PIC)

- Founded in 2003
  - Collaboration between IFAE and CIEMAT
- Supports lots of projects:
  - Spanish Tier-1 WLCG: Atlas, CMS, LHCb
  - Astro: MAGIC, PAU, MICE, DES, ...
- Resources:
  - 6000 cores, 10 PiB disk, 14 PiB tape

- Euclid:
  - SDC-ES
  - Primary SDC for OU-SIM
    - Integrate, run and debug simulation pipelines
  - CS-SWG: Flagship mock galaxy catalog

Flagship data regime

- Flagship halo catalog
  - ~5.5 TiB
  - ~40 B haloes
- Flagship mock galaxy catalog (full sky, no cuts)
  - ~20 TiB
  - **~60 B galaxies**
- Needs special infrastructure
  - We had some previous experience with Big Data
    - Our choice was clear: Hadoop

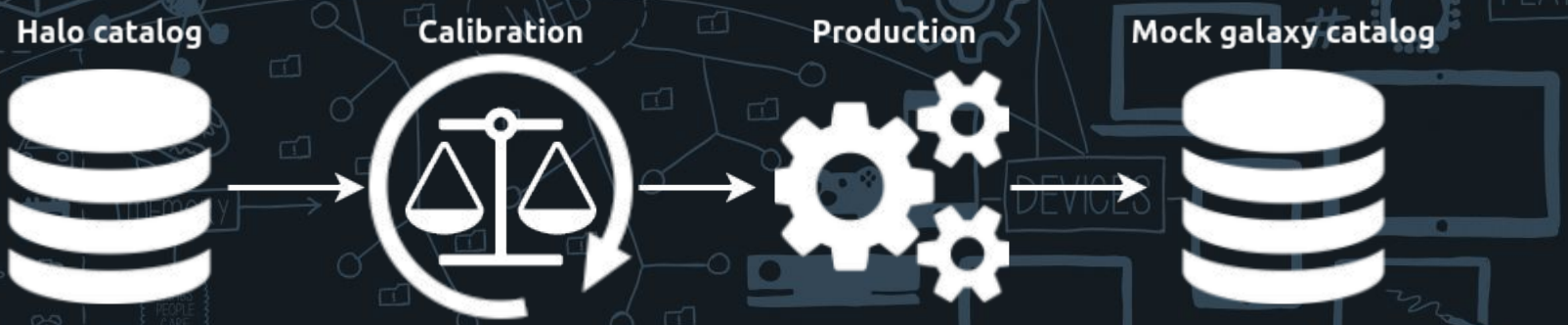
PIC Big Data platform

- Based on Hadoop (Hortonworks HDP 2.5)
  - Open source
  - Distributed storage and processing
    - Runs on commodity computer clusters
    - Scalable from dozens up to thousands of nodes
      - ✓ Performance scales with HW
    - Failure tolerance
      - ✓ May use old/refurbished machines → cheaper

- PIC Hadoop platform
  - 30 nodes: 360 cores, 1080 GiB RAM, 90 TiB HDD

SciPIC

- Scientific pipeline at PIC
- Set of Python codes/algorithms to compute:
  - HOD
  - Galaxy properties:
    - Luminosities, positions, colors, velocities, sed, stellar properties, emission lines, magnitudes, morphology
  - Clustering



SciPIC performance

	50 deg²	400 deg² (SPV)	octant (H<26)
# input halos	47.9 M	386 M	5.08 B
# output galaxies	69.6 M	561 M	2.63 B
size on disk	42 GiB	179 GiB	0.83 TiB
time spent	8.1 min	39 min	5.3 h

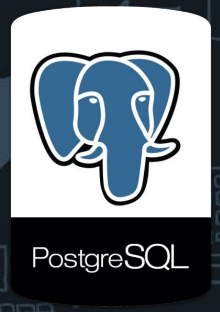
COSMO HUB

- Main features:
  - Generate and download custom catalogs
    - Guided process, no SQL knowledge needed
    - Expert mode available
  - Interactive visualization
  - Value-added-data ready to download
- Datasets:
  - Euclid, DES, Gaia, MICE, PAU, SDSS, COSMOS, USNO
- CosmoHub in numbers:
  - ~ 400 users, > 1500 custom catalogs, > 5000 plots

Nov 2016: Rewrite and port on top of Hadoop

COSMO HUB before Big Data

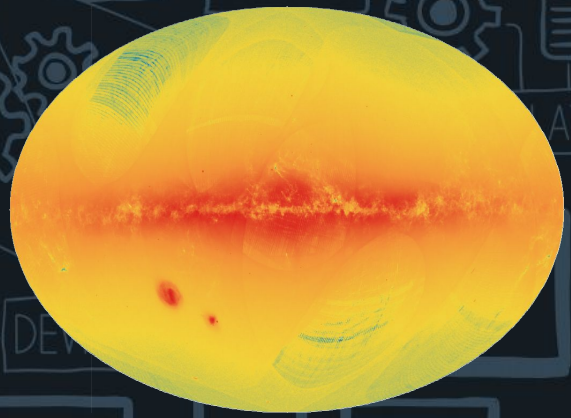
- Storage
  - Based on PostgreSQL relational database
  - Limited to server HDD space (30 TiB)
- Performance
  - Does not scale as data grows
  - Query time range: minutes to **days**
- Visualization
  - Limited to 10.000 rows or 2 minutes
  - Becomes meaningless as data grows

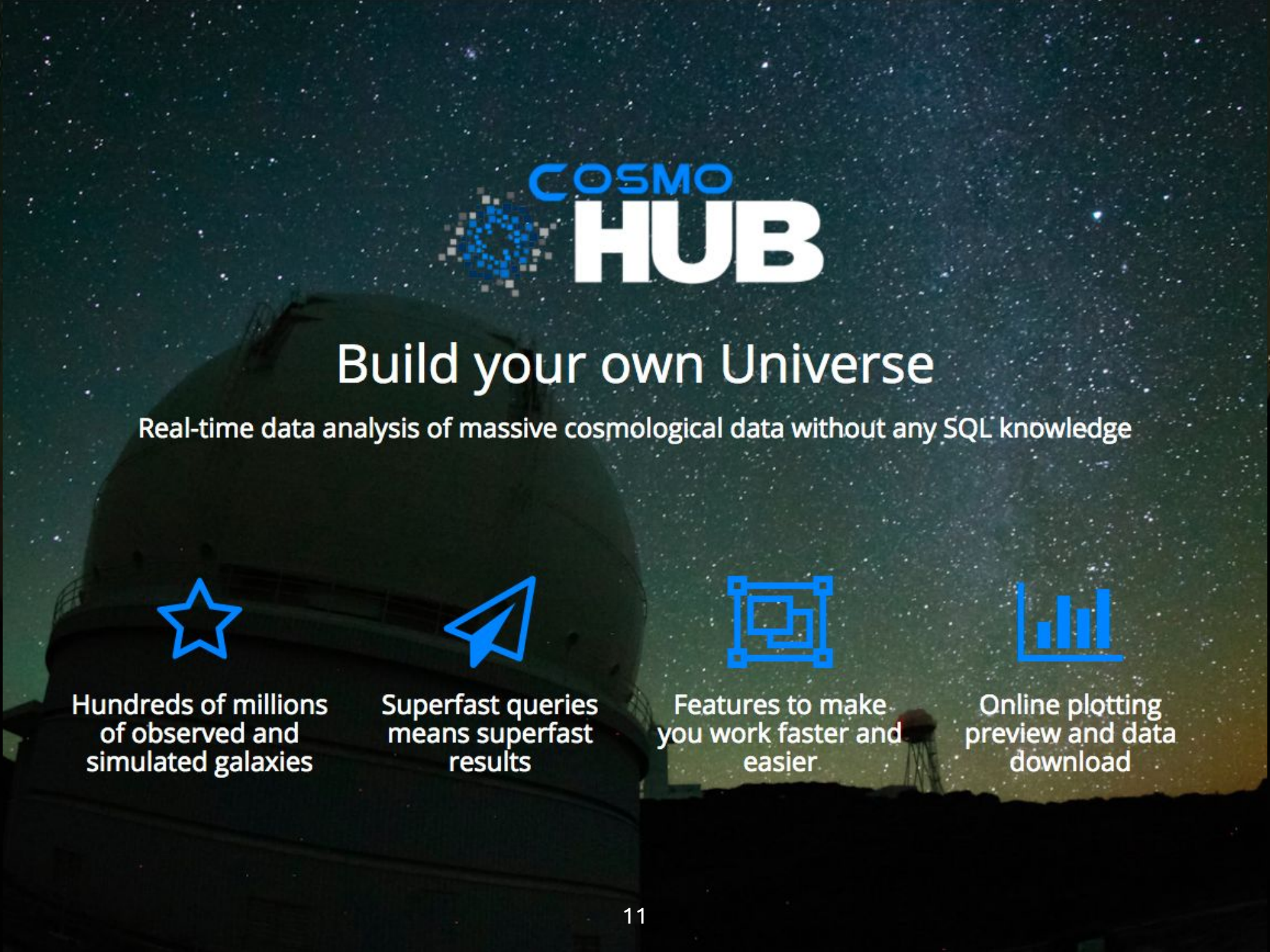


Is NOT PostgreSQL fault!

COSMO HUB after Big Data

- Storage
  - Based on Hadoop (HDFS) + Hive
  - Scales with HW (90 TiB and growing)
- Performance
  - Huge improvement **(x100 speedup)**
  - Query time range: seconds to minutes*
    - 85% in < 3 min
- Visualization
  - Unlimited time
  - Full dataset plots (over all rows)
    - May use sampling
    - 1D histogram & 2D heatmap




COSMO HUB
Build your own Universe
Real-time data analysis of massive cosmological data without any SQL knowledge
Hundreds of millions of observed and simulated galaxies
Superfast queries means superfast results
Features to make you work faster and easier
Online plotting preview and data download

Conclusions & Future work

- Hadoop has proven to be a great choice
- SciPIC
  - Production and calibration of galaxy mocks
    - Very fast -> many iterations -> better results
  - WIP: integrate lensing and photo-z
- CosmoHub
  - HUGE performance improvement over Hadoop
  - Switched perspective
    - from catalog generation to interactive analysis
  - Stay tuned for next upgrades ;)

Team - Q&A

- CosmoHub
  Jorge Carretero, Jordi Casals, Marc Caubet, Santiago Serrano, Pau Tallada

- SciPIC
  Alex Alarcón, Linda Blot, Jorge Carretero, Francisco Castander, Pablo Fosalba, Santiago Serrano, Pau Tallada

- Special thanks to
  Carles Acosta, Jordi Delgado, Davide Piscia, Nadia Tonello, Francesc Torradeflot and the rest of PIC staff