Big Data @ NERSC

- Data sharing and analytic Services

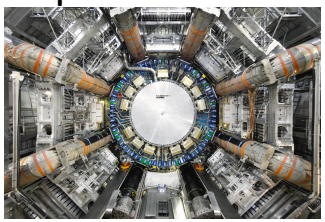
Yushu Yao

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Uses of Data at NERSC

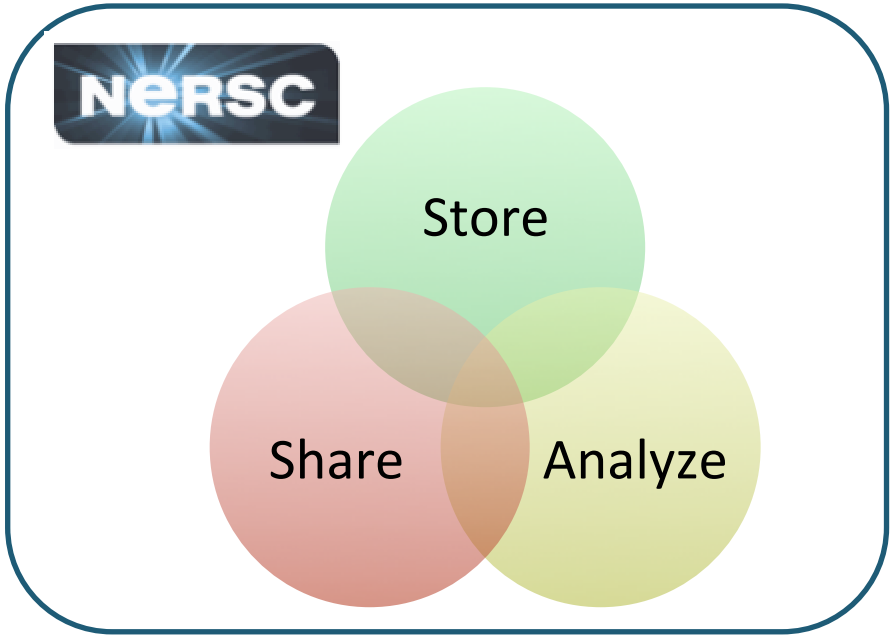
Experiment



Computer Simulation

Data

Data



- **Data comes to (or generated at) NERSC from Apparatus or Computer Simulations**
- **Three Things People Do:**

Store Share Analyze

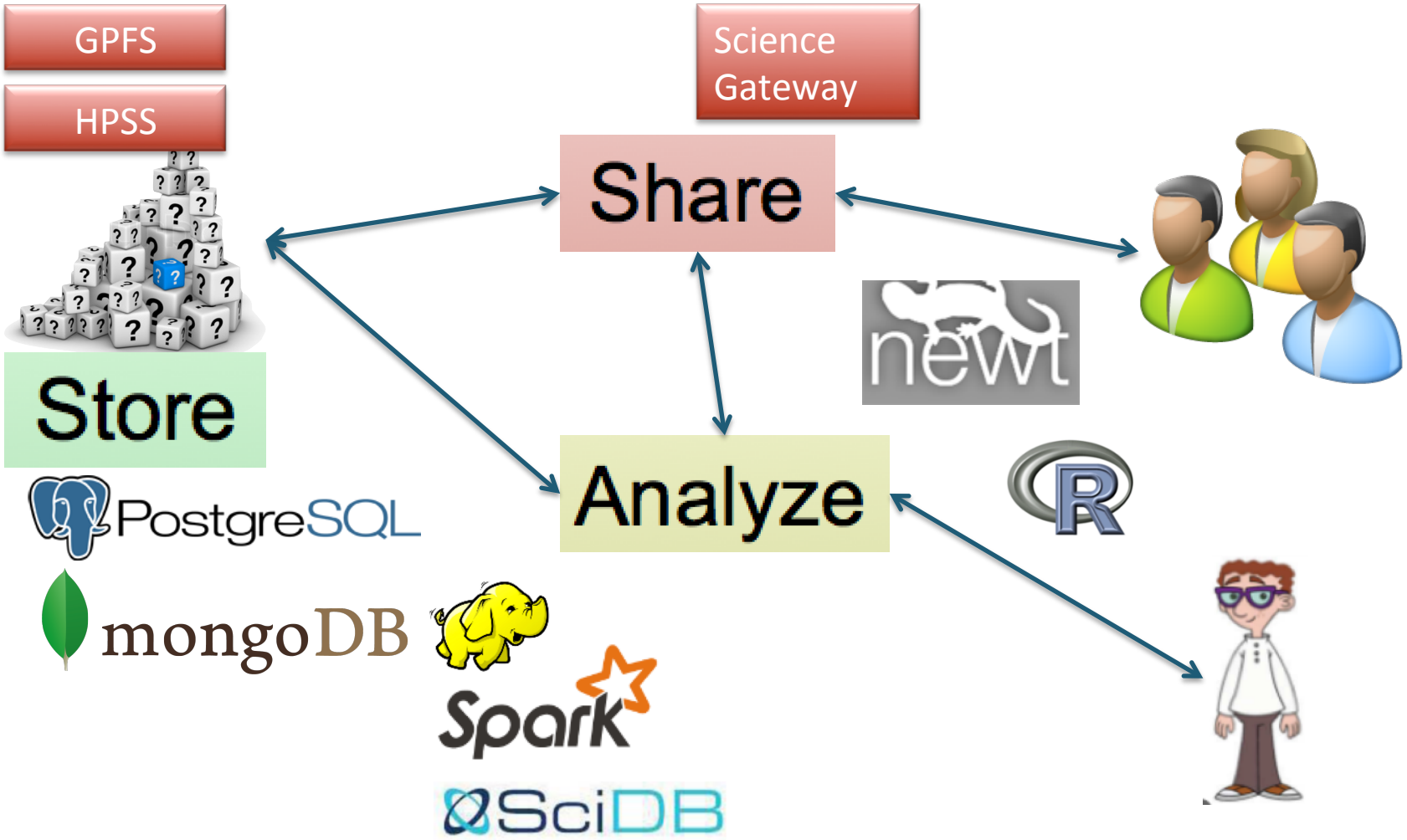
Store/Share/Analyze Data At NERSC

Store

Share

Analyze

Store/Share/Analyze Data At NERSC



Science Gateway Services

- **Publish data on the web**
  - Create a www directory in your project space and put your data on the web
- **Build sophisticated web portals**
  - Build full stack web applications for your science at NERSC using Python/Django, PHP, Ruby on Rails

- **NEWT – the NERSC REST API**
  - Use the NEWT REST HTTP API to access NERSC HPC resources directly from your web apps.
  - Support for Authentication, Jobs, Commands, Files, Persistent Store, NIM, System Information at NERSC

Gateway examples - http://portal.nersc.gov
NEWT – https://newt.nersc.gov
Science Gateway information http://www.nersc.gov/users/science-gateways/

Traditional SQL Database Services

- **PostgreSQL and MySQL**

- **Good For:**
  - Structured, Relational Data
  - Mid-Size, <=several GB in total
  - Transactional Operations

MongoDB

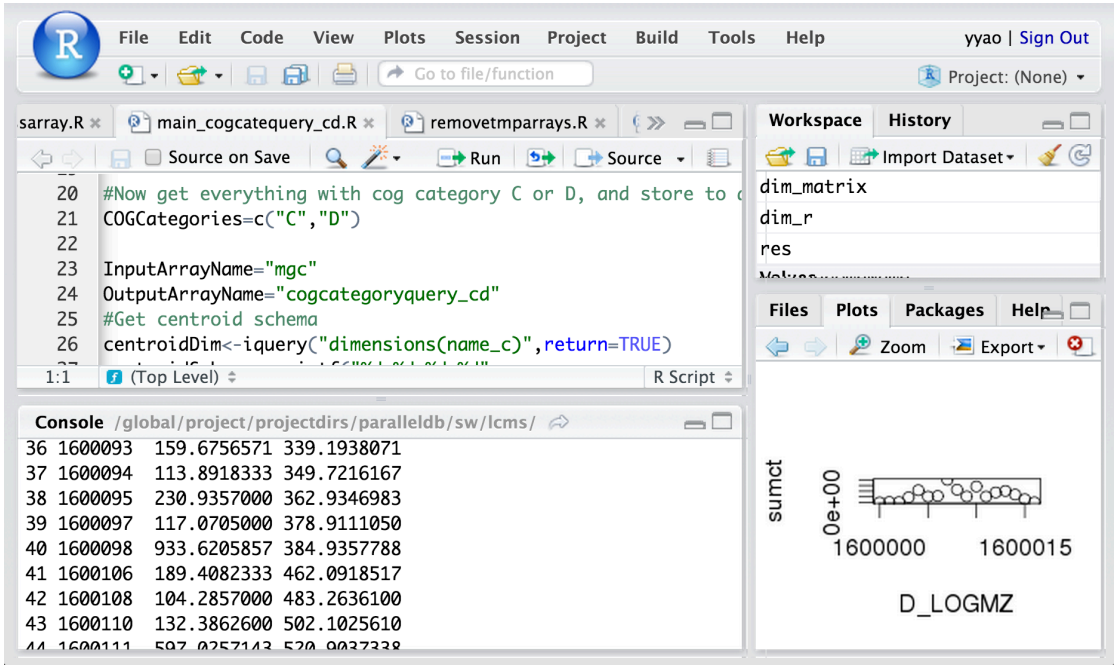
- **Key-value pair / Text database**

- **Good For:**
  - Un-Structured, Text Data
  - Mid-Size to Large, e.g. 10 GB of Text
  - E.g: for metadata that has ever changing schema

To request a MongoDB/PostgreSql/MySQL database:
https://www.nersc.gov/users/science-gateways/science-gateway-databases/

RStudio Beta Service (r.nersc.gov)

- **Simply go to http://r.nersc.gov and login with NIM username/password**
- **IDE for the popular R analytic package:**
  - Access to Global File Systems (Project, etc)
  - Rich set of machine learning and other analytic packages
  - Web-based, pick up unfinished work on any other browser.



Spark/Hadoop Testbed

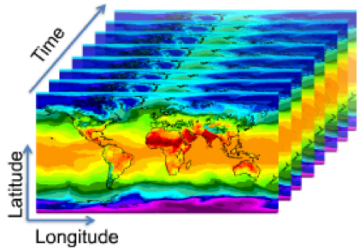
- **Map Reduce on Large Amount of Data**

- **Good For:**
  - Un-Structured or structured, not easily indexed
  - Large, e.g. 100 GB or more.
  - E.g: Large amount of sequencing data

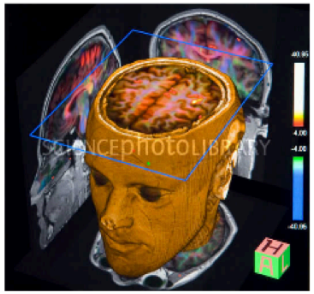
SciDB Testbed: Array-Like Data Examples

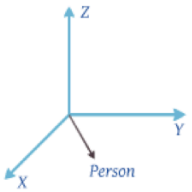
Climate Simulation Output

-Terabytes of Output Per Run



Brain MRI Image

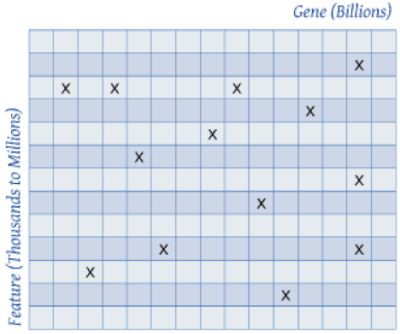




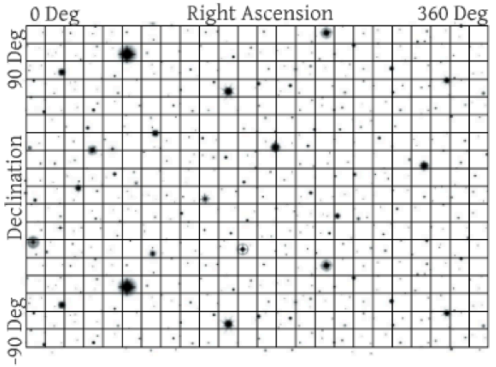
Many of people -> 4th Dimension

Gene Labeling

- Large Sparse Array



Catalog of Billions of Stars



SciDB Testbed at NERSC

- **Large Parallel Array Database/Analytic**
- **Good For:**
  - Large (10GB~10TB) array structure data

- *Partner up with Science Teams*
  - *Hold their hands to load 1st batch data, do 1st batch of analysis.*
- *10+ Science Projects*
- *Complicated Algorithms*
- *Multiple Science Domains:*
  - *Astronomy, Climate, Bio-imaging, Genomic*

More Information

- **For more information about the SciDB and Spark testbeds at NERSC, please email consult@nersc.gov**

National Energy Research Scientific Computing Center