Databases

Wahid Bhimji
(wbhimji@lbl.gov)

https://www.nersc.gov/users/data-analytics/data-management/
# NERSC Data and Analytics Stack

## Capabilities
- Processing
- Storage/Management
- Analytics/Visualisation
- Access
- Transfer

## Tools
- Fireworks/Swift
- python/R/ROOT
- OMERO/Fiji/Matlab
- Visit/Paraview
- NEWT
- GridFtp

## Services
- BDAS/SPARK
- iPython/RStudio
- HDF5/NetCDF
- SciDB/MongoDB
- Postgres/MySQL
- Globus

## Systems
- Compute Nodes
- Interactive Nodes
- Burst Buffer
- Parallel Filesystem
- Global FS
- Database Servers
- Science Gateways
- Data Transfer Nodes
Databases

Capabilities

- Storage/Management
- Analytics/Visualisation
- Access

I/O Libraries and Scientific I/O

- HDF5/NetCDF
- SciDB/MongoDB
- Postgres/MySQL

Databases

Services

- Burst Buffer
- Parallel Filesystem

Systems

- Database Servers
Databases

• Relational / SQL Databases
  – NERSC host/support MySQL and PostgreSQL DBs for users

• NoSQL / Schema-less Databases
  – MongoDB

• Array Databases
  – SciDB

To request a database:
https://www.nersc.gov/users/science-gateways/science-gateway-databases/
• **Good for:**
  – Your data is structured (you have a ‘Schema’)
  – Relational (tables of rows and columns)
  – Mid-Size, <= 50 GB in total
  – transactional operations (ensuring DB is consistent)

• Single databases on single servers so multiple connections not served in parallel.

• **PostgreSQL:** Object relational, some powerful features and extensions (including geospatial and astro libraries) as well as SQL standards

• **MySQL:** Very popular, open-source, relational database
Accessing SQL DBs

- **Postgres:**
  
  ```bash
  psql -h scidb1.nersc.gov yourdb -U dbuser
  ```

- **Mysql:**
  
  ```bash
  mysql yourdb -u dbuser -h scidb1.nersc.gov -p
  ```

- **Very basic sql:**
  
  ```sql
  SET PASSWORD = PASSWORD('password');
  USE yourdb;
  CREATE TABLE yourtable (a_id INTEGER PRIMARY KEY ,b VARCHAR(10) );
  SELECT * from yourtable WHERE yourtable.b = 'bob';
  ```

- **Learning SQL try**  
  
  [http://sqlzoo.net/](http://sqlzoo.net/)
• ‘NoSQL’, document-oriented database
• JSON-like documents (key: value)
• Queries are javascript expressions
• Memory-mapped files – queries can be fast
• Good For:
  – Un-Structured Data (‘Schema-less’)
  – Mid-Size to Large, <= ~ 2-3 TB
• Though not configured for very frequent/ high-volume writes or very many connections
  – More than 1 insert /s ask us / try test MongoDB instance
Accessing MongoDB

• Use mongo client
  
mongo -u yourdb_admin -p password mongodb01.nersc.gov/yourdb

• Create a collection; put a document in it and find it
  
doc1 = {name: “bob”, friends:5 }
  
yourdb.acollection.insert(doc1)
  
db.customers.find({name:"bob"})

• Use pymongo for Python:
  
  import pymongo
  
  client= pymongo.MongoClient(‘mongodb01.nersc.gov’) 
  
  client.admin.authenticate(yourdb_admin, args.passwd)
  
  client.yourdb.acollection.insert([{"name": ”bob”, ”friends”: 5}])
SciDB

- ‘Array’ Database: Time series / Spectra / Images etc.
- Scalable to big data (TBs+) – ‘horizontally’
- Perform statistical analyses in parallel without HPC programming using queries

Good For:
  - Array Structured Data (Dense or Sparse)
  - Large, >10+ GB up to 10+ TB
  - You want to push array type computations on big data and return statistical summaries (not extract the big data)

Boutique service – contact us via web form if interested:

https://www.nersc.gov/users/science-gateways/science-gateway-databases/
Accessing databases (and data)

• Can use command line or APIs
• Or via Webpages – e.g. Science Gateways or Omero:
  • Visualization, management and analysis of (biological microscope) images
    – CLI (‘dropbox’ folder) to upload images
    – Read any image format
    – Connect via web portal or query...
    – Perform operations (averages, max, min etc.)
• Request PostgreSQL via web form and specify Omero in comments
Parting thoughts on databases

- **Files and Databases (some personal observations)**
  - Massively parallel HPC programs -> Files
  - Instrument data distributed around the world -> Files
  - Large 100-1000 user collaborations -> Files
  - metadata (e.g. about conditions in which data was collected) -> Database (SQL if schema known)
  - Multi-source, aggregated, instrument metadata -> NoSQL DB
  - Array analytics -> SciDB

- **Database and file I/O documentation:**

- **Database Request Form**