Databases and Data Analytic Frameworks at NERSC

Yushu Yao (yyao@lbl.gov)

July 15, 2014
NERSC Data Analytic Services

Big and Diverse Computing Facility
6000+ Users, 700+ Projects
3+ PetaFlops (20+pf more coming)
50+ PB Storage
Store/Share/Analyze Data At NERSC

Store

Share

Analyze

Yushu Yao
Store/Share/Analyze Data At NERSC

GPFS
HPSS

Store

Science
Gateway

Share

PostgreSQL
mongoDB

Analyze

Spark
SciDB
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](http://portal.nersc.gov)
NEWT – [https://newt.nersc.gov](https://newt.nersc.gov)
Example Gateway: Materials Project

http://www.materialsproject.org

- Web Server Hosted at NERSC
- Jobs Submitted to NERSC Systems as needed
- Data Stored in different DB and Storage Systems
- You can do this too
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/
SciDB For High Usability Big Data Analytic

• **Why?** It’s painful to manage and analyze terabytes of data. Need a unified solution that’s easy to use.

• **What?** SciDB is a parallel database for array-structured data, great for **Terabytes** of:
  – Time series, spectrums, imaging, etc

• The greatest benefit of SciDB is:
  – **Usability**: Use HPC hardware without learning parallel programming and parallel I/O.
Spark for In-Memory Map-Reduce

Spark
Lightning-Fast Cluster Computing


• Very efficient in-memory map-reduce data analytic framework.
Any NERSC user can go to r.nersc.gov and login with NIM Username/Password
**iPython Notebook Service (Coming Soon)**

**IPython Notebook**

**Lecture_3_Scipy** (autosaved)

---

**In [6]:**

```python
x = linspace(0, 10, 100)

fig, ax = subplots()
for n in range(4):
    ax.plot(x, jn(n, x), label=r'J_n(x)')
ax.legend();
```

---

**In [7]:**

```python
# zeros of Bessel functions
n = 0  # order
m = 4  # number of roots to compute
jn_zeros(n, m)
```

**Out[7]:**

```
array([ 2.40482556,  5.52007811,  8.65372791, 11.79153444])
```
NERSC Data Analytic Services

Big and Diverse Computing Facility
6000+ Users, 700+ Projects
3+ PetaFlops (20+pf more coming)
50+ PB Storage
Links

- **Science Gateways:**

- **Database Request Form**

- **SciDB**

- **Rstudio**
  - [http://r.nersc.gov](http://r.nersc.gov)

- **Spark**