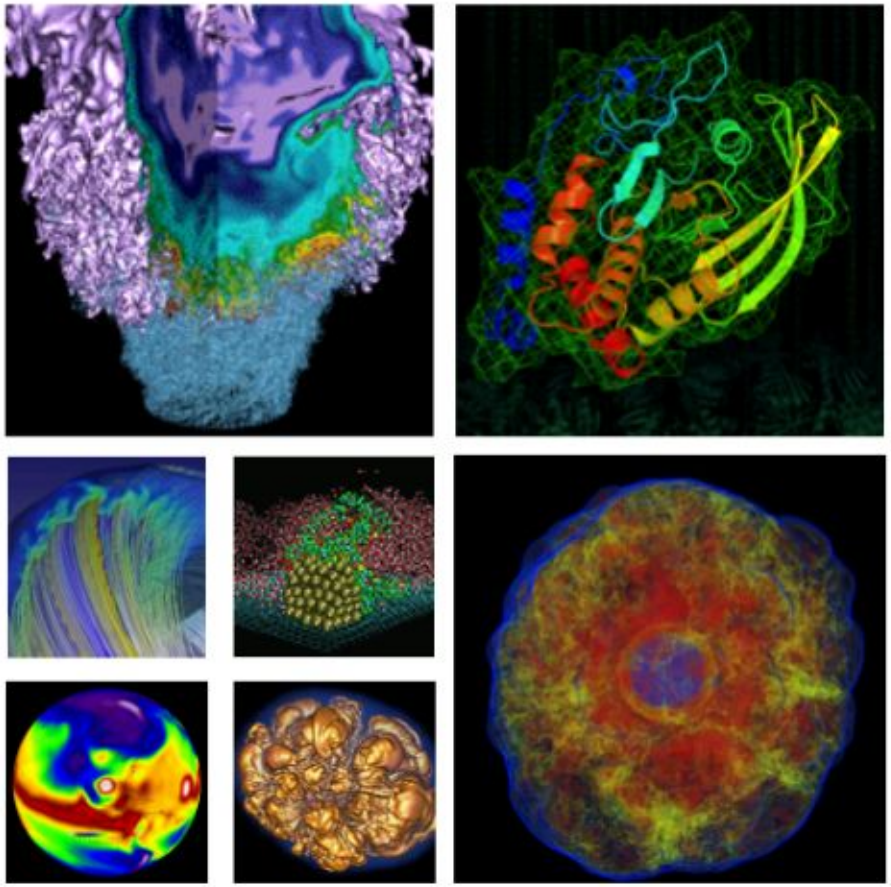
Using Spark on Cori



Lisa Gerhardt, Evan Racah
NERSC New User Training

U.S. DEPARTMENT OF ENERGY

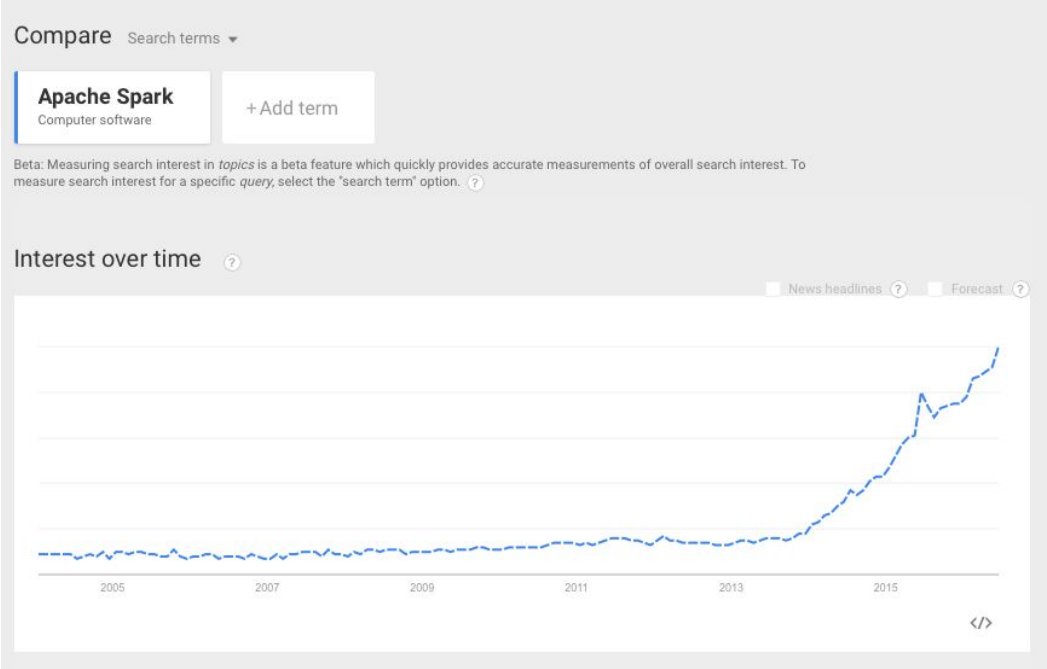
BERKELEY LAB

What is Spark?

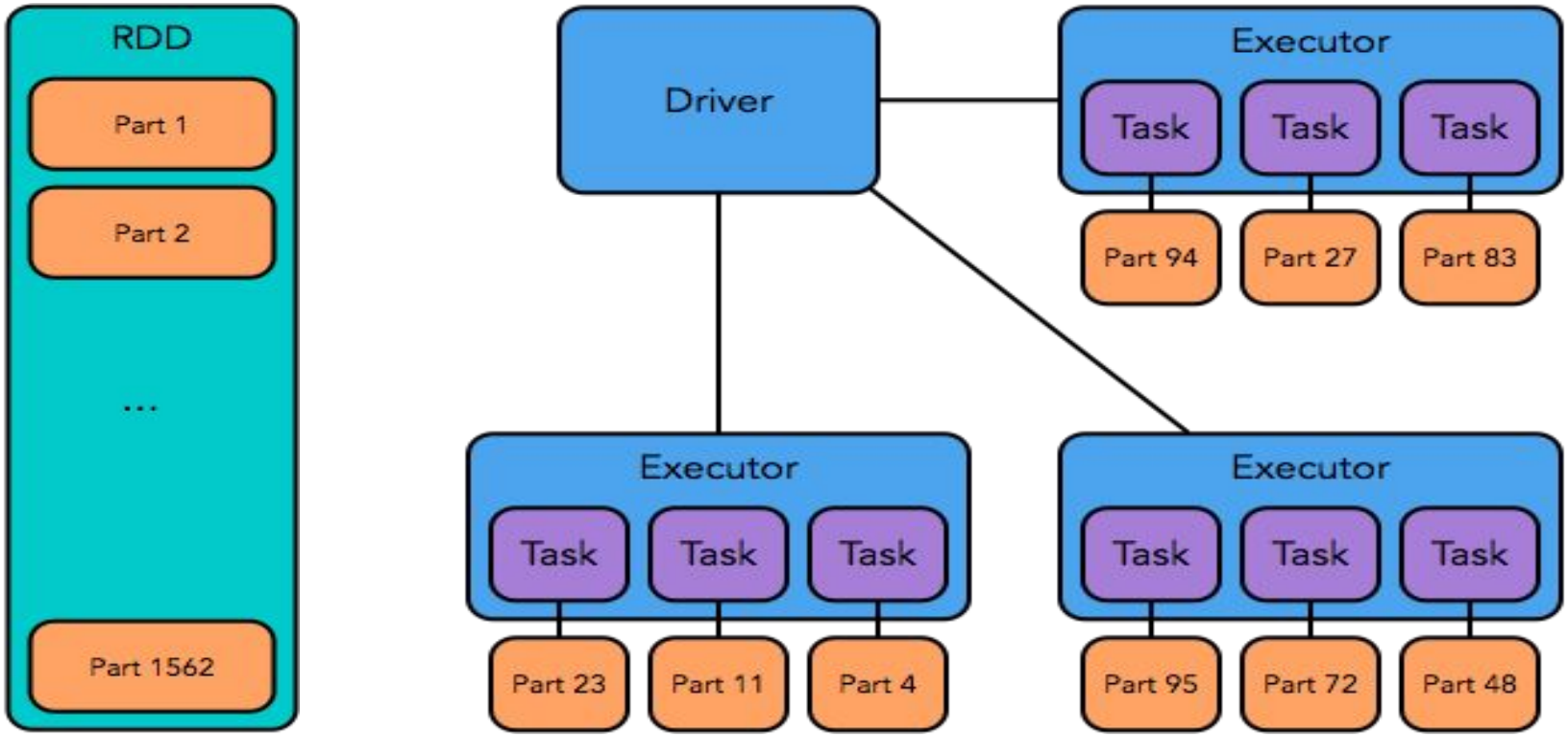
- Apache Spark is a fast and general engine for large-scale data processing
- **Data-parallel model**
  - Similar to Hadoop, except it harnesses in-memory data for fast data processing
- Robust body of libraries: machine learning, sql, graphx, streaming
- Created at the AMPLab in UCB, currently maintained by DataBricks, more than 400 contributors to git repository
- **Requires no parallel programming expertise**: Can take Python, Java, or Scala code and automatically parallelize it to an arbitrary number of nodes

Spark: So hot right now

- Spark Summit in June had more than 2500 attendees
- Used by Airbnb, amazon, eBay, Groupon, IBM, MyFitnessPal, NetFlix, OpenTable, Salesforce, Uber, etc.
- And by NASA (SETI), UCB (ADAM)
- Used at NERSC for metagenome analysis, astronomy, and business intelligence



Spark Architecture



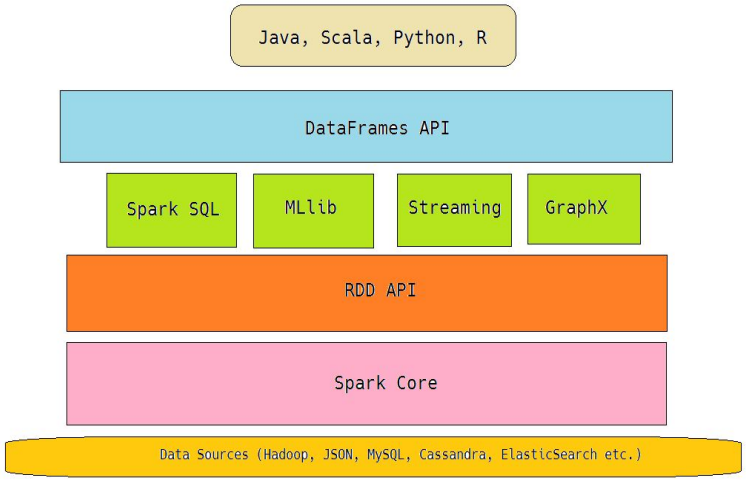
- One driver process, many executors
- Driver creates DAG and coordinates data and communication (via akka TCP)
- Datasets generally cached in memory, but can spill to local disk

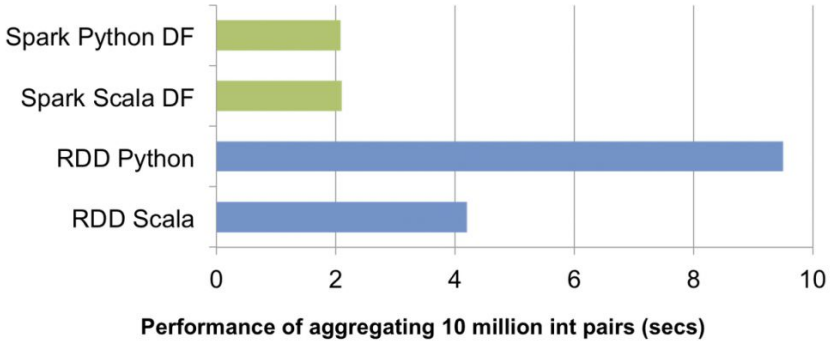
Spark API

- Can use Java, Python, or Scala
- Latest version also has an R interface and a Jupyter notebook
- Most newer functionality comes to Scala interface first
- Also some indications that Scala performs better
- Python is easier to use, so we recommend using that to start out
  - Can always reoptimize in scala later if necessary
- Default version at NERSC is Spark 2.0.0
- Recommend running on Cori
  - Larger memory nodes and faster connections with scratch file system

Spark Data Formats

- **RDD**: Resilient Distributed Dataset. Immutable collection of **unstructured** data.
- **DataFrames**: same as RDD but for **structured** data. Familiar pandas-like interface
- Which to use?
  - Future: DataFrames
  - Now: DataFrames and fall back to RDDs for more fine tuned control





Spark Goodies

- **Streaming**: high-throughput, fault-tolerant stream processing of live data. Twitter, kafka, etc.
- **SQL**: Structured data processing, lives on top of RDD API
- **ML**: classification, regression, clustering, collaborative filtering, feature extraction, transformation, dimensionality reduction, and selection. Best only for really LARGE datasets.
- **GraphX**: Graph parallel computation. Extends RDD to a directed multigraph with properties attached to each vertex and edge

Spark Strategy

- **Best for DATA PARALLEL!!!**
- **Make sure you have enough memory**
  - Datasets are held in memory for speed
  - When memory is filled up, must spill data to disk, which slows things considerably
    - NERSC default spills to RAM file system and to Lustre scratch
  - If you get “java.lang.OutOfMemoryError” your executors have run out of memory. Fix by making the load on each node smaller. Either
    - Increase the number of nodes **OR**
    - Increase executor memory
- **Keep the cores fed**
  - Optimal number of partitions is 2 - 3 times the number of cores, automatic partitioning sizes can be a little wonky

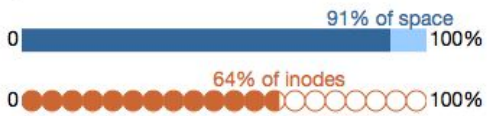
Real World Use Case: Data Dashboard

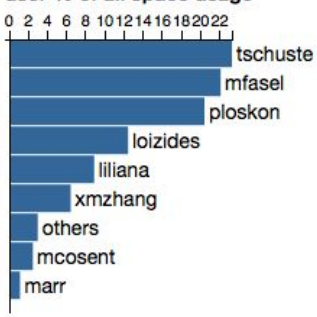
Data Dashboard

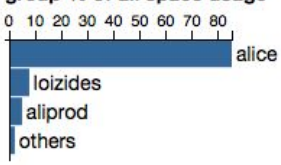
Showing disk space and inode usage for project directories at NERSC to which you have access as PI, PI proxy, or user (includes /project, /projecta, and /projectb/sandbox)

alice directory in **/project** as of Mon Jun 27 2016 04:45:04 GMT-0700 (PDT) Toggle Details

Usage as Percent of Allocations







On MyNERSC
https://my.nersc.gov/data-mgt.php

```
3 1 666  131072 root 1 D2u 12 0 0 drwxr-xr-x 1467007091 1465523705 131072 1465524073 -- %2Fproject%2F.snapshots%2F2016-06-26
6 2020878165 666  327 dayabay 2020878165 FAu 1 44768 41120 -rw-r--r-- 1419249645 1419249646 4194304 1419249646 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fdayabay%2Fwww%2Fdybprod%2Flogs%2Fruns_0048000%2Fru
7 65542 666  2202398 cosmo 65542 FAu 1 58191 58191 -rw-r--r-- 1456894476 1305484507 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F1b%2F01701
8 65544 666  23383 cosmo 65544 FAu 1 58191 58191 -rw-r--r-- 1453037873 1305485593 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F8a%2F01698a
9 65544 666  181440 des 65544 FAu 1 57947 56799 -rw-r----- 1368137836 1367293101 4194304 1367293101 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fdes%2Fwl%2Fdesdata%2Fusers%2Frarmst%2Ffixed_weight3%2FDECam_
10 65547 666  68519 cosmo 65547 FAu 1 58191 58191 -rw-r--r-- 1456767186 1305483688 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F1a%2F01701
11 65543 666  4170240 cosmo 65543 FAu 1 58191 58191 -rw-r--r-- 1452963922 1305484507 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F1b%2F0170
12 65546 666  2060841 cosmo 65546 FAu 1 58191 58191 -rw-r--r-- 1453037873 1305485593 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F8a%2F0169
13 65542 666  2211647 cosmo 65542 FAu 1 58191 58191 -rw-r--r-- 1456767186 1305483686 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F1a%2F0170
14 65542 666  1074240 cosmo 65542 FAu 1 58191 58191 -rw-r--r-- 1453037873 1305485593 4194304 1371829105 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fcosmo%2Fdata%2Fwise%2Fallsky%2F4band_p1bm_frm%2F8a%2F0169
15 65538 666  9935 dayabay 65538 FAu 1 44768 44768 -rwxrwxrwx 1348676160 1348676160 4194304 1353297189 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fdayabay%2Fwww%2Fdybprod%2Fdiagnostics%2Fruns_0033000%2Fru
16 65538 666  10861 dayabay 65538 FAu 1 44768 44768 -rwxrwxrwx 1348676160 1348676160 4194304 1353297189 -- %2Fproject%2F.snapshots%2F2016-06-26%2Fprojectdirs%2Fdayabay%2Fwww%2Fdybprod%2Fdiagnostics%2Fruns_0033000%2Fru
```

200 GB of text

Demo

https://github.com/NERSC/data-day-examples

Conclusion

- Spark is an excellent tool for data parallel tasks that are too big to fit on a single node
  - Easy to use python DataFrames interface
  - Does parallel heavy lifting for you
- Spark 2.0 installed at NERSC, performs well on data-friendly Cori system

National Energy Research Scientific Computing Center

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB
Lawrence Berkeley National Laboratory

Spark Batch Job

```
#!/bin/bash

#SBATCH -p debug
#SBATCH -N 2
#SBATCH -t 00:30:00
#SBATCH -e mysparkjob_%j.err
#SBATCH -o mysparkjob_%j.out
#SBATCH --ccm

module load spark
start-all.sh
spark-submit --master $SPARKURL --driver-memory 15G --executor-memory
20G ./testspark.py
stop-all.sh
```

Submit with sbatch <submitscriptname.slurm>

Spark Terminology

- **Driver process**: runs on head node and coordinates workers (you can adjust memory allocation)
- **Executor**: single process that runs on worker nodes
- **Task**: a unit of work that's sent to one executor
- **Application**: entire program that is run
- **Job**: spawned by a spark action (like *collect*), consists of one or more stages on multiple nodes
- **Stage**: a set of tasks that do the same operation on a different slice of the data
  - See frequent references to "job" and "stage" in the spark logs
- **RDD** (Resilient Distributed Datasets): read only collection of records, held in memory or spilled out to disk