

# Data Management Part 3

Yusra AlSayyad  
LSST Data Management  
Princeton University

DSFP- Session 3  
April 28 2017





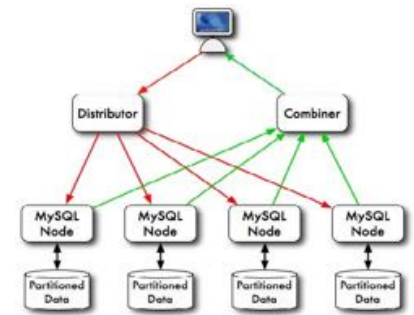
# Outline

- While Spark Demo is running:
  - Hardware (cloud?)
  - Spark
  - Used in Industry?

# Hardware



Extreme Science and Engineering  
Discovery Environment



## Demo



- Student accounts come with free \$ for AWS:  
<https://aws.amazon.com/education/awseducate/>
- <https://aws.amazon.com/emr/pricing/>
- Start cluster at: [console.aws.amazon.com](https://console.aws.amazon.com)

# Gartner's Hype Cycle

2012





# 2014



2016



Years to mainstream adoption:

○ less than 2 years


○ 2 to 5 years

● 5 to 10 years

▲ more than 10 years

obscure

■ before plateau



“But what’s happening is that big data has quickly moved over the Peak of Inflated Expectations...and has become prevalent in our lives across many hype cycles.”

“I would not consider big data to be an emerging technology”

Betsy Burton



# MapReduce-based systems NOT a replacement for:

- HPC cluster computing
- Databases!
  - Transactional vs Analytic
    - OLTP vs. OLAP
  - Latency
  - CAP Theorem

Example: Eventually consistent key-value store. highly available



If data is partitioned,  
Choose between:



Availability

Consistency

# Spark

	Hadoop Map Reduce	Spark
Storage	Disk Only	In-memory or disk
Execution	Batch	Batch, interactive, streaming
Language	Java	Java, python, R, scala

- Won Daytona-Gray sorting contest 2014
  - 100 TB in 23 minutes
  - Details: 207 Amazon EC2 i2.8xlarge nodes x (32 vCores - 2.5Ghz Intel Xeon E5-2670 v2, 244GB memory, 8x800 GB SSD)



## Spark in Industry

- For Netflix:
  - 25-40% of analytics jobs, shifting away from Pig/Hive.
  - Like it for ETL (Extract, Transform, and Load),
- Still seen as an immature technology:
  - Cryptic error messages
  - Performance/Tuning problems



## Thoughts

- If you need “Big Data” tools, you would know: existing tools wouldn’t work for your problem
  - While it has promise, your existing workflow will probably not be improved by running it in Spark (unlike other topics we covered this week)
  - When you are responsible buying hardware for your research group, consider cloud as an option
- 