


# Jason Geng

Founder and CEO at Data Application Lab

University of Southern California • Texas A&M University

Greater Los Angeles Area • 500+

**JasonGeng@DataAppLab.com**

**https://www.linkedin.com/in/gengjason/**

Data Application Lab

# Big Data Analytics Fundamental

Jason Geng

jason@DataAppLab.com

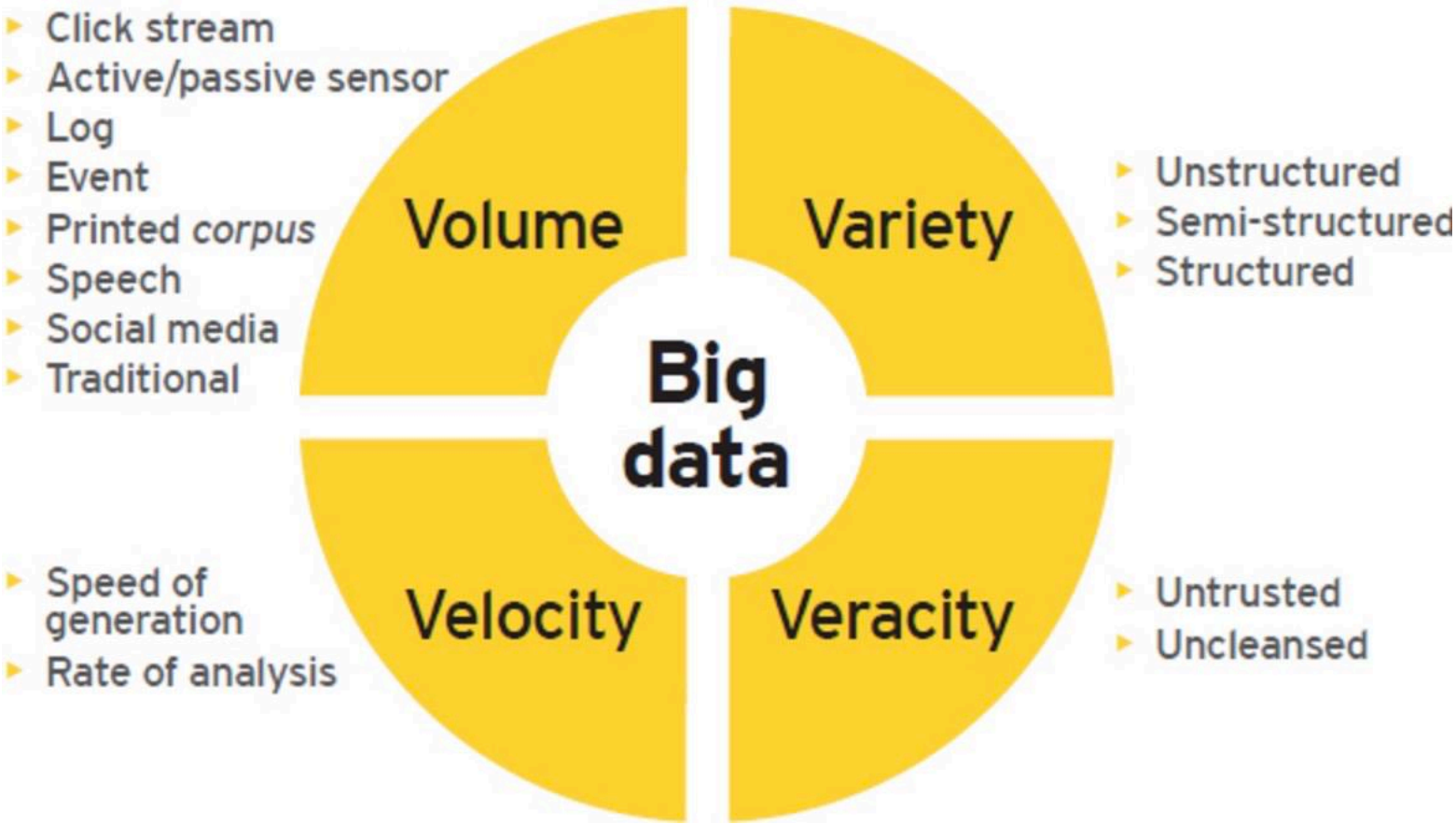
Data Application Lab

# Disclaimer

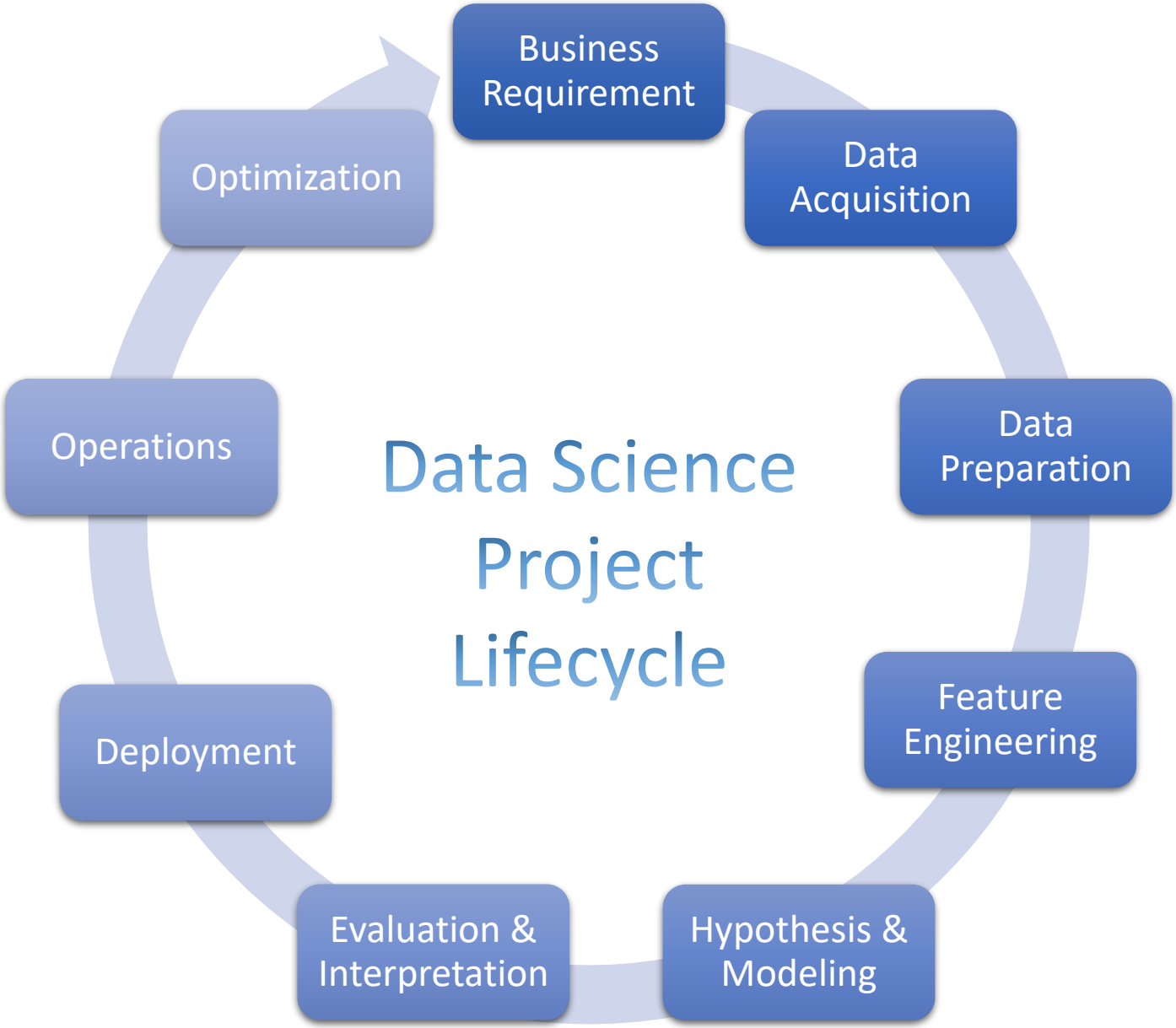
- All data and information in this presentation were from public resource
- This is a vendor-independent talk that expresses teacher own opinions
- 

Data Application Lab

# What is Big Data




Data Application Lab




Data Application Lab

# Business Requirements

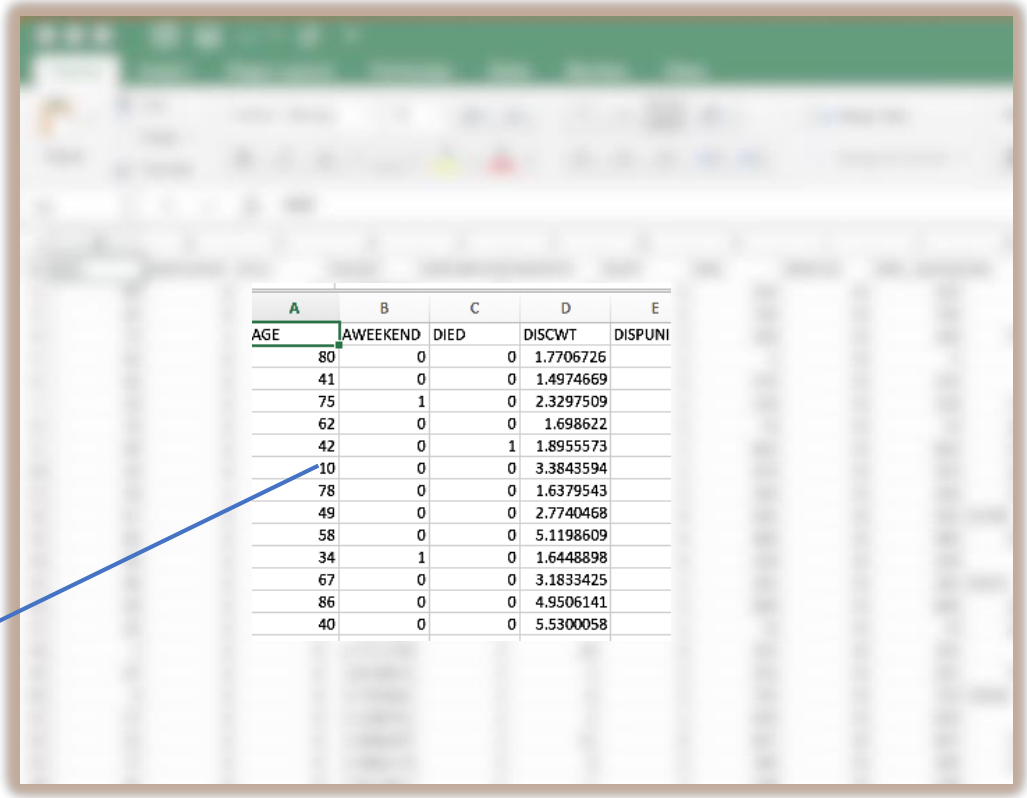
- Data scientists need to work with business people and those with expertise in understanding the data, understanding the business
- Specify the business requirements
- For instance, the healthcare data

- Database:

**Healthcare:**
Readmissions Database

- Understand the data:

**e.g. 'DISCWT':**
'This the discharge-level weight on the HCUP nationwide data to produce national estimates'



| A | B | C | D | E |
|---|---|---|---|---|
| AGE | AWEEKEND | DIED | DISCWT | DISPUNI |
| 80 | 0 | 0 | 1.7706726 | |
| 41 | 0 | 0 | 1.4974669 | |
| 75 | 1 | 0 | 2.3297509 | |
| 62 | 0 | 0 | 1.698622 | |
| 42 | 0 | 1 | 1.8955573 | |
| 10 | 0 | 0 | 3.3843594 | |
| 78 | 0 | 0 | 1.6379543 | |
| 49 | 0 | 0 | 2.7740468 | |
| 58 | 0 | 0 | 5.1198609 | |
| 34 | 1 | 0 | 1.6448898 | |
| 67 | 0 | 0 | 3.1833425 | |
| 86 | 0 | 0 | 4.9506141 | |
| 40 | 0 | 0 | 5.5300058 | |

- Understand the Business:

**Goal:**
Predict Readmission Rate

→ Feature Exploring → Modeling

Data Application Lab

## Database:

User

Session

Purchase

Usage

## Understand the data:

e.g. 'IAP':
'In App Purchase

| A | B | C | D | E |
|---|---|---|---|---|
| AGE | AWEEKEND | DIED | DISCWT | DISPUNI |
| 80 | 0 | 0 | 1.7706726 | |
| 41 | 0 | 0 | 1.4974669 | |
| 75 | 1 | 0 | 2.3297509 | |
| 62 | 0 | 0 | 1.698622 | |
| 42 | 0 | 1 | 1.8955573 | |
| 10 | 0 | 0 | 3.3843594 | |
| 78 | 0 | 0 | 1.6379543 | |
| 49 | 0 | 0 | 2.7740468 | |
| 58 | 0 | 0 | 5.1198609 | |
| 34 | 1 | 0 | 1.6448898 | |
| 67 | 0 | 0 | 3.1833425 | |
| 86 | 0 | 0 | 4.9506141 | |
| 40 | 0 | 0 | 5.5300058 | |

## Business Requirements:

Goal:
Increase "Non-pay" player purchases

Exploring

Solution

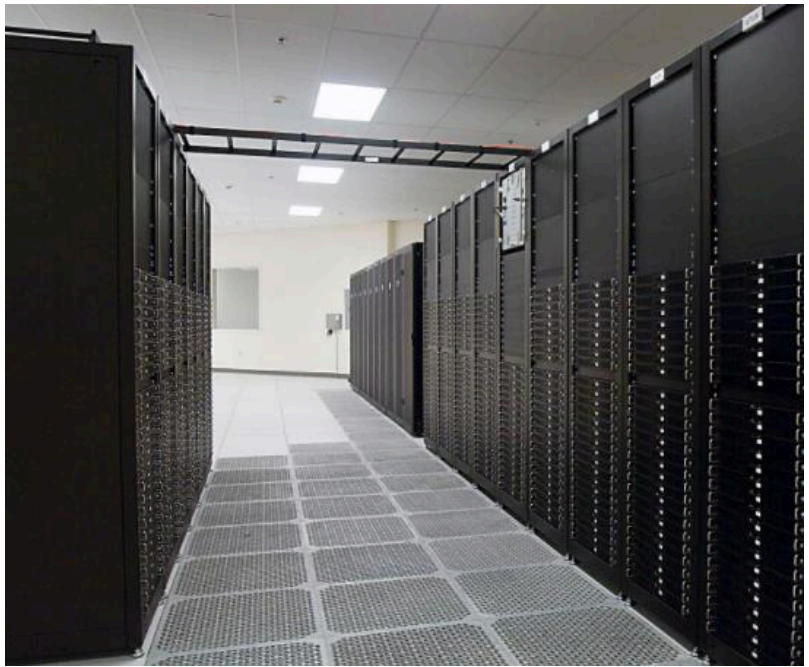
Data Application Lab

# Data Collection

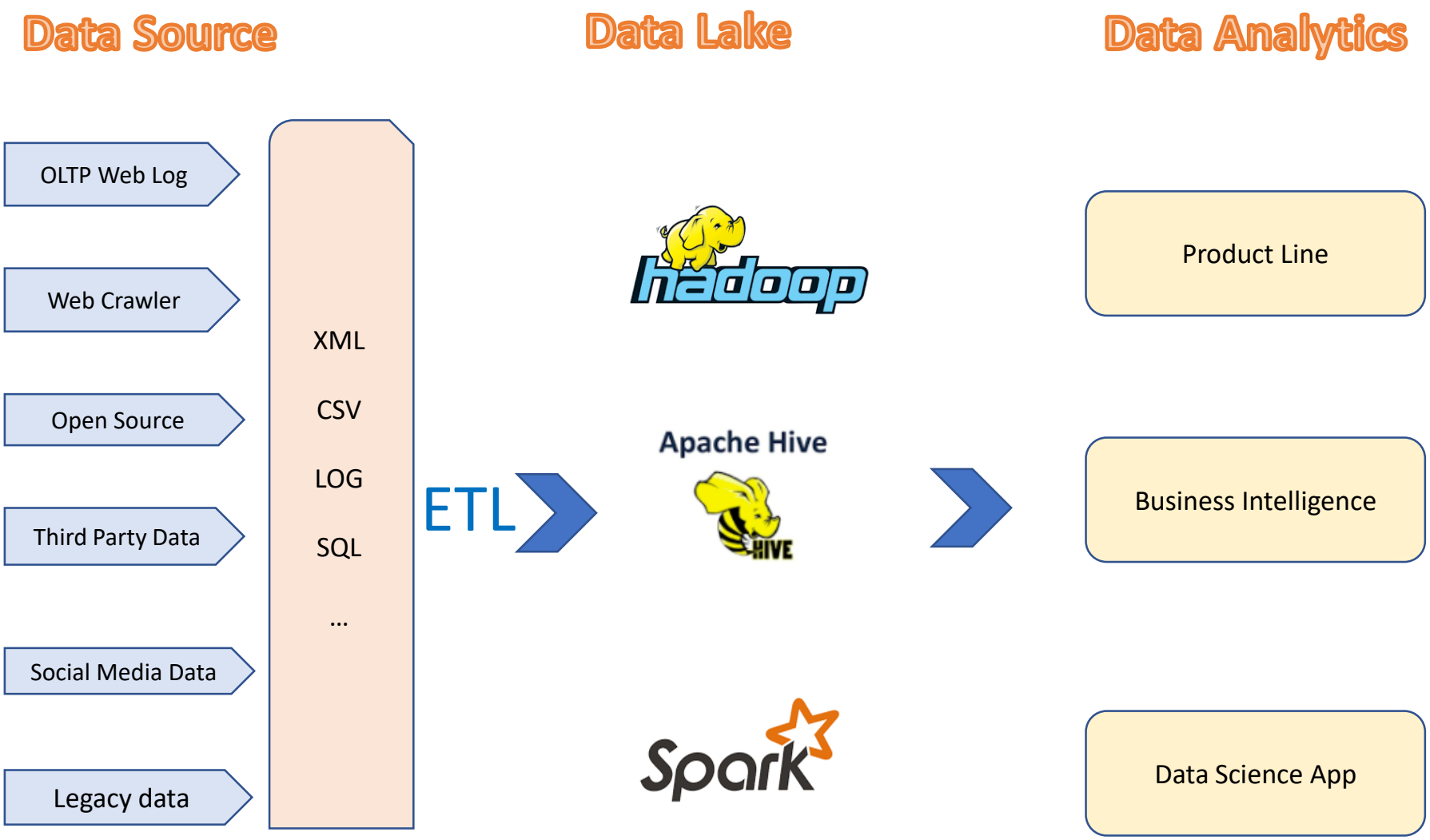
- Data from product line
- Purchase third party data
- Social media (Facebook, LinkedIn)
- Web crawling
- Open source data (Opendata, U.S. Census Data)

Challenge

- Data Storage
- Data Management



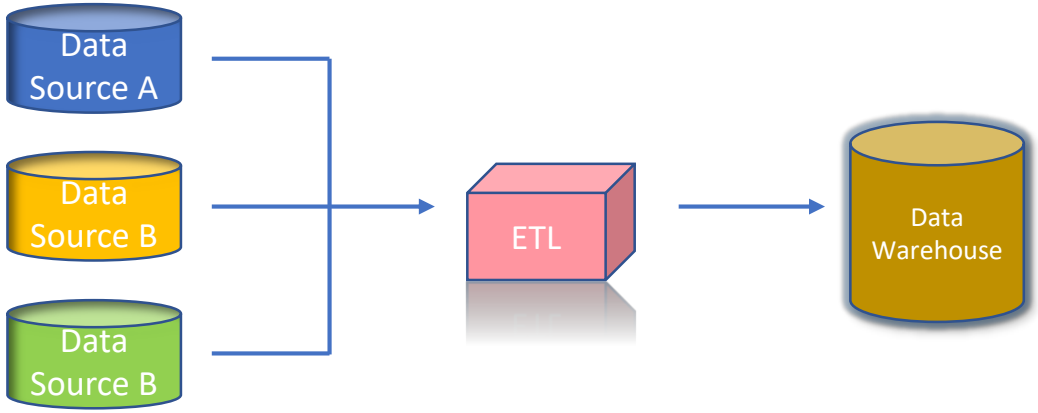
Data Application Lab




Data Application Lab

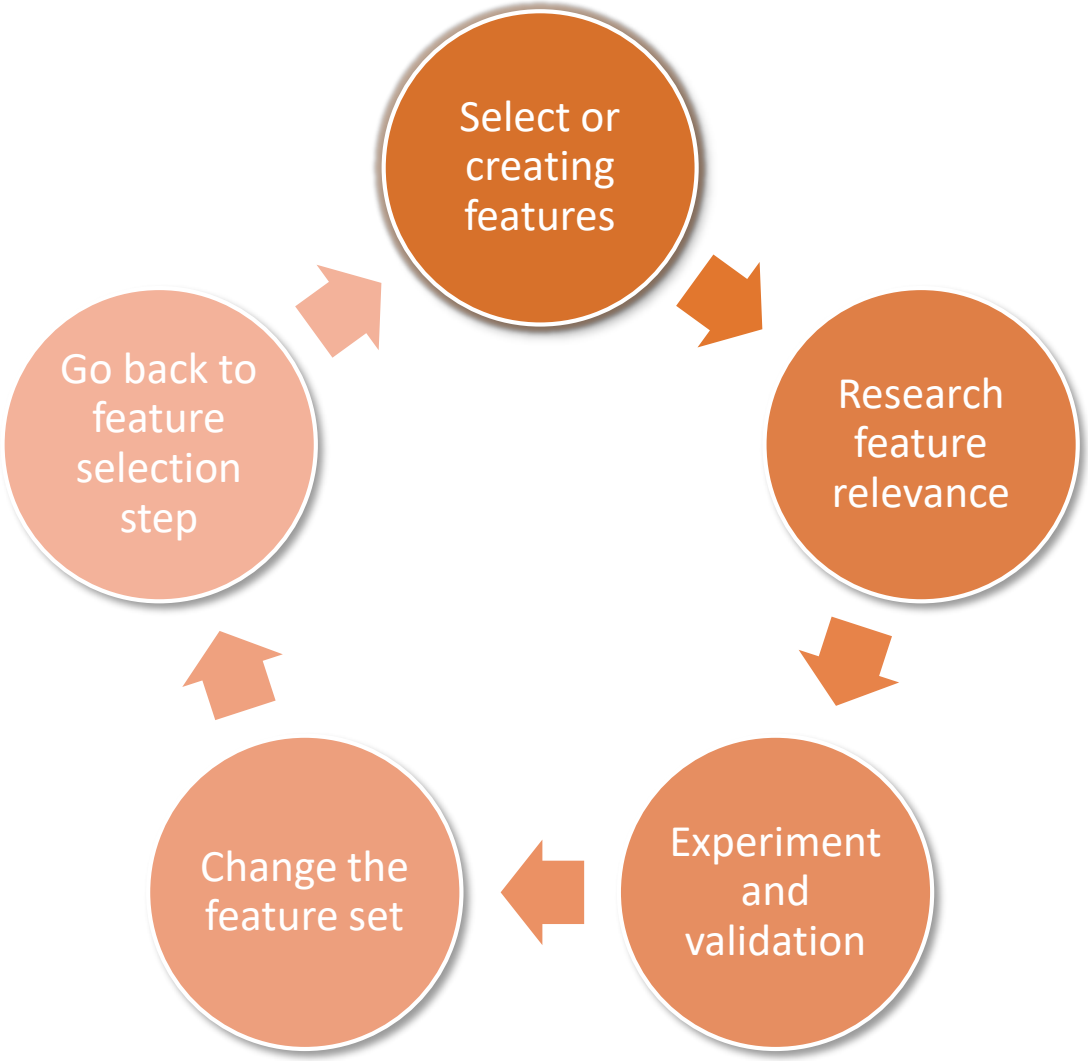
# Data Preparation (Data Wrangling)

- Cleaning data (semantic errors, missing entries, or inconsistent formatting)
- Challenge: data integration
- **80%** time in project workflow



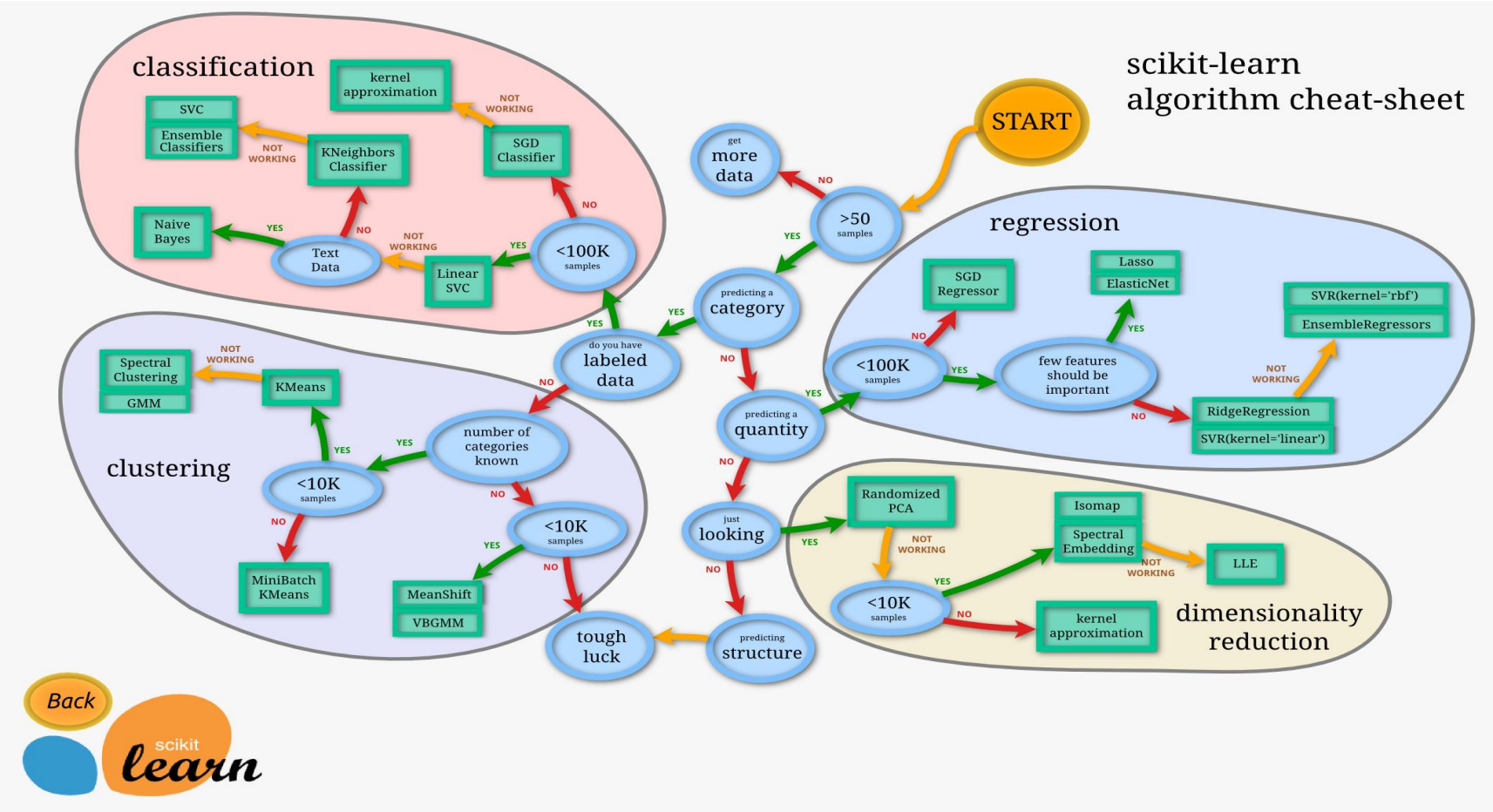

Data Application Lab

# Feature Engineering




Data Application Lab

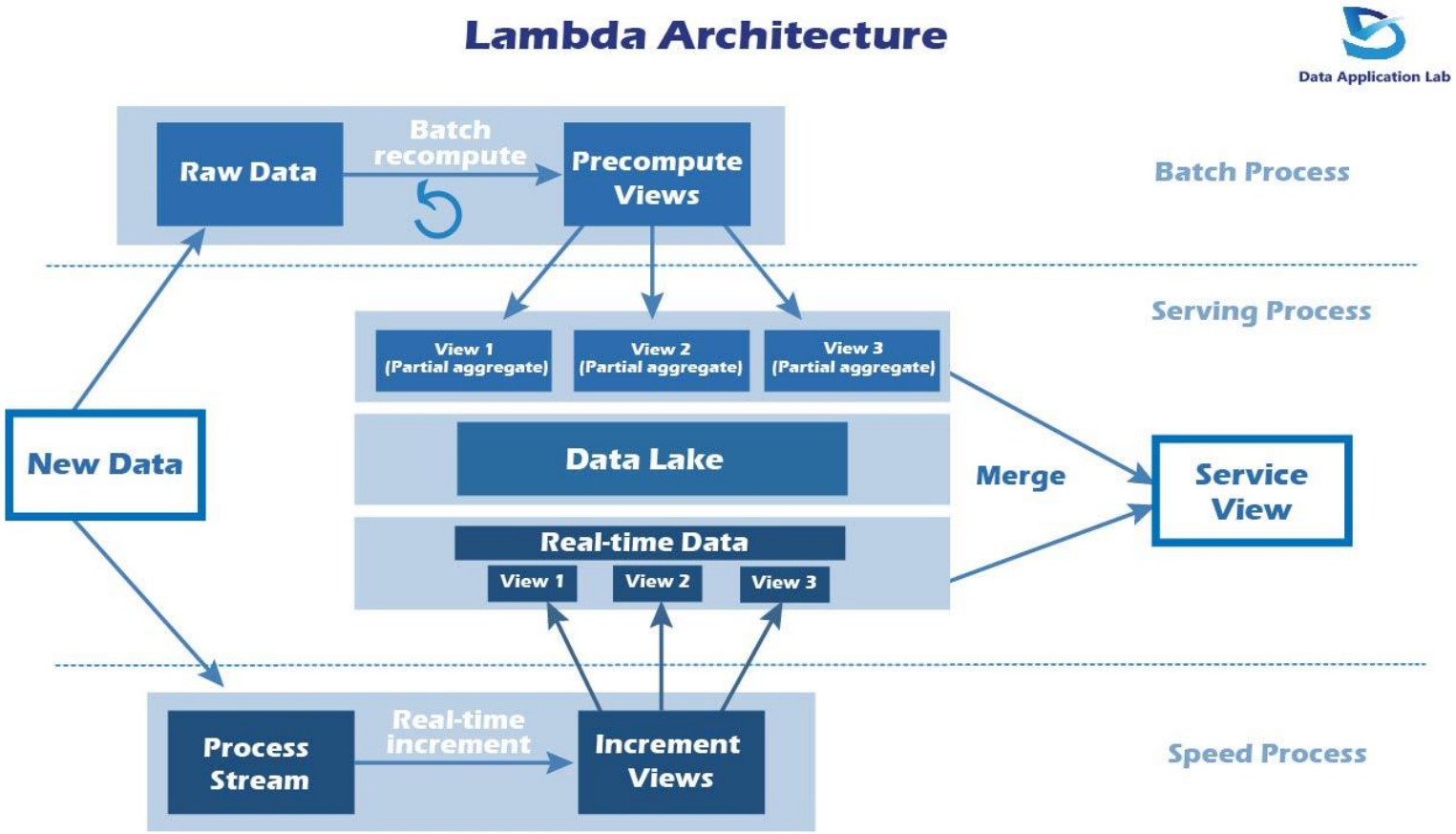
# Modeling




*Reference Source: http://scikit-learn.org/stable/tutorial/machine_learning_map/*

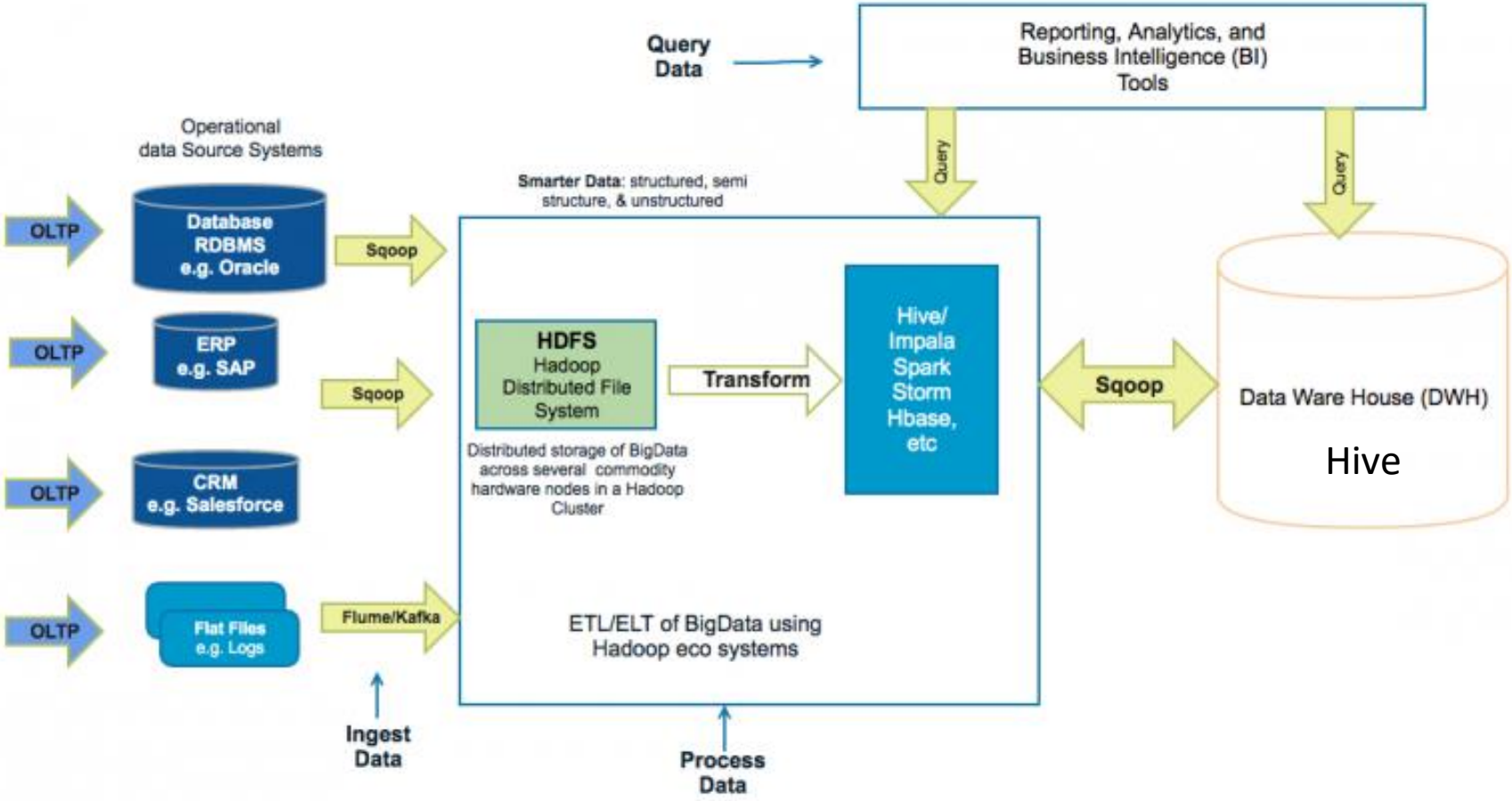
Data Application Lab

# Deploy to Product Line




Data Application Lab

# Hadoop Data Warehouse




Data Application Lab

# Computer cluster

From Wikipedia, the free encyclopedia

*Not to be confused with data cluster or computer lab.*
*"Cluster computing" redirects here. For the journal, see Cluster Computing (journal).*

A **computer cluster** consists of a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system. Unlike grid computers, computer clusters have each node set to perform the same task, controlled and scheduled by software.[1][*better source needed*]

The components of a cluster are usually connected to each other through fast local area networks ("LAN"), with each *node* (computer used as a server) running its own instance of an operating system. In most circumstances, all of the nodes use the same hardware[2] and the same operating system, although in some setups (i.e. using Open Source Cluster Application Resources (OSCAR)), different operating systems can be used on each computer, and/or different hardware.[3]

They are usually deployed to improve performance and availability over that of a single computer, while typically being much more cost-effective than single computers of comparable speed or availability.[4]

Computer clusters emerged as a result of convergence of a number of computing trends including the availability of low-cost microprocessors, high speed networks, and software for high-performance distributed computing.[*citation needed*] They have a wide range of applicability and deployment, ranging from small business clusters with a handful of nodes to some of the fastest supercomputers in the world such as IBM's Sequoia.[5]

Technicians working on a large Linux cluster at the Chemnitz University of Technology, Germany

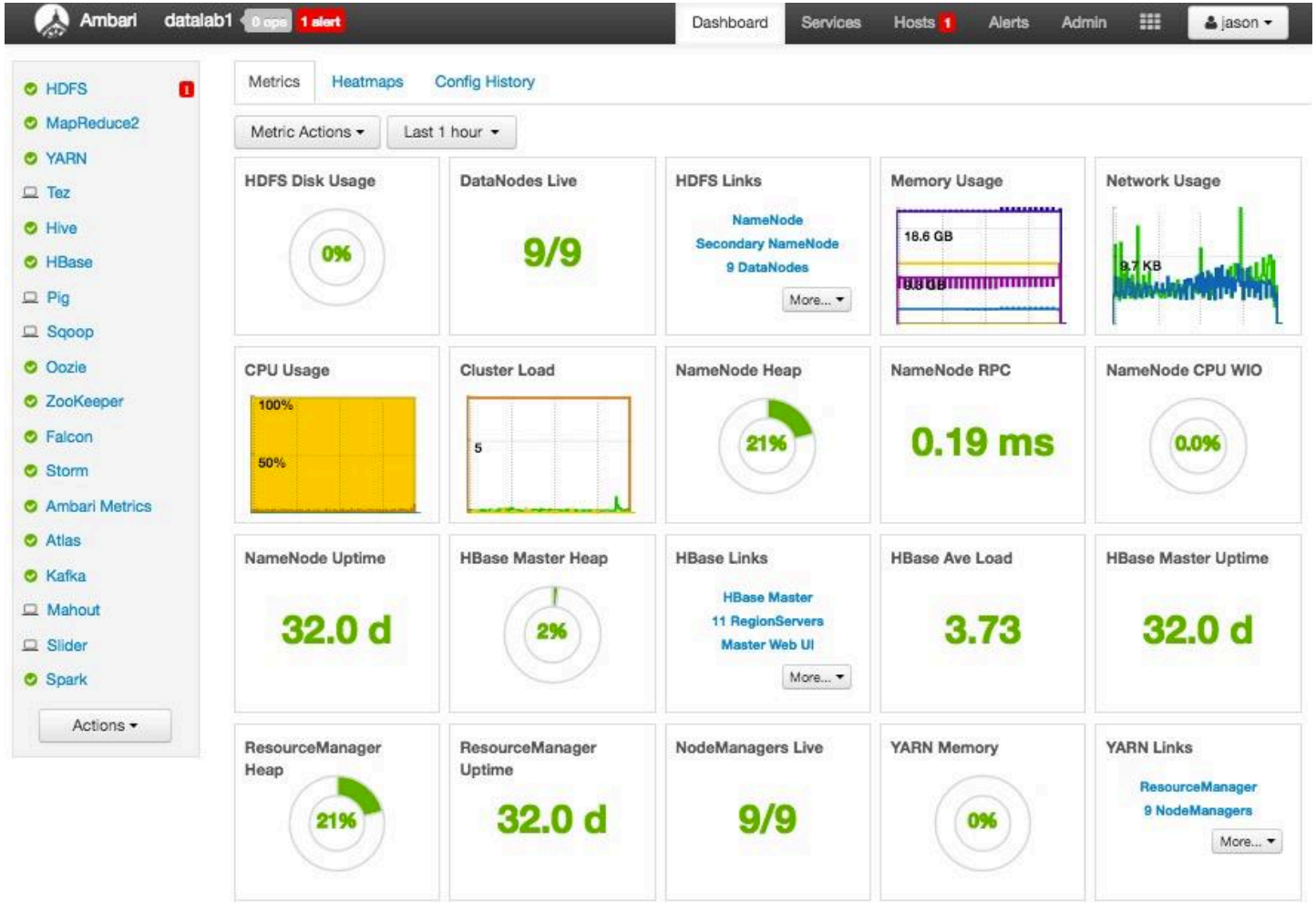

Sun Microsystems Solaris Cluster

**Contents** [hide]



Data Application Lab

# DAL Cluster




Data Application Lab

How to install putty on Windows
https://www.youtube.com/watch?v=a4K9mvKxrwI

how to use winscp
https://www.youtube.com/watch?v=e7AgOFS_g8Q

how to use SSH on Mac
https://www.youtube.com/watch?v=J_8ZsXP1EYk

how to use scp on mac
https://www.youtube.com/watch?v=EJOoiYtyPTE

VI tutorial
https://www.youtube.com/watch?v=TBu6qxd5uAc

# HDFS Design Goal

- Recover from hardware failure
- Streaming data access
- Large data file/dataset
- Write-once-read-many IO model
- **Move computation to data**
- Commodity hardware
- Do not conflict with OS file system
- **Not good** – low-latency and many small files

# From local machine to remote machine

Login

ssh jason@montana.dataapplab.com -p 49233

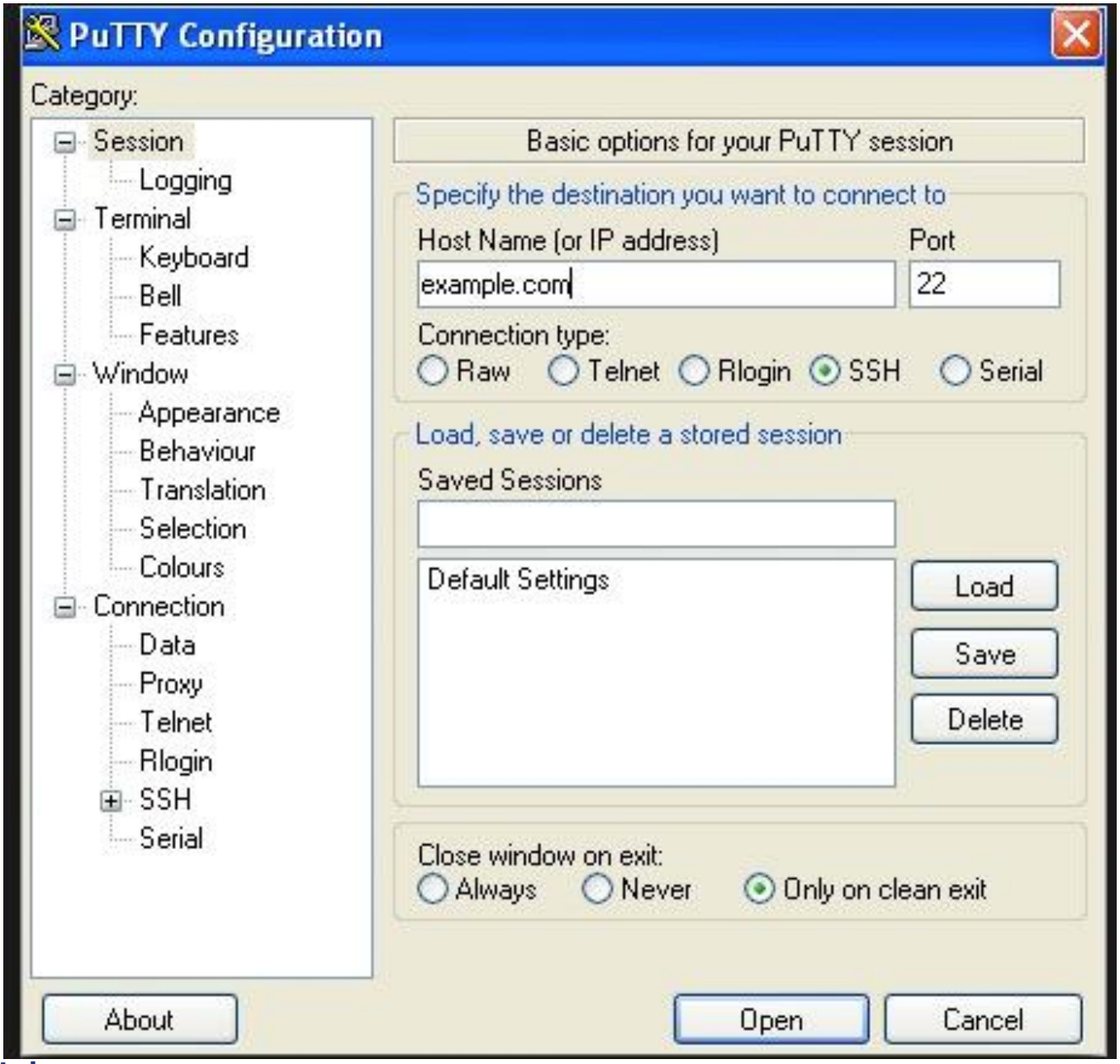
Copy files

scp -P 49233 /Users/jason/folder/file jason@montana.dataapplab.com :/home/useraccount/folder

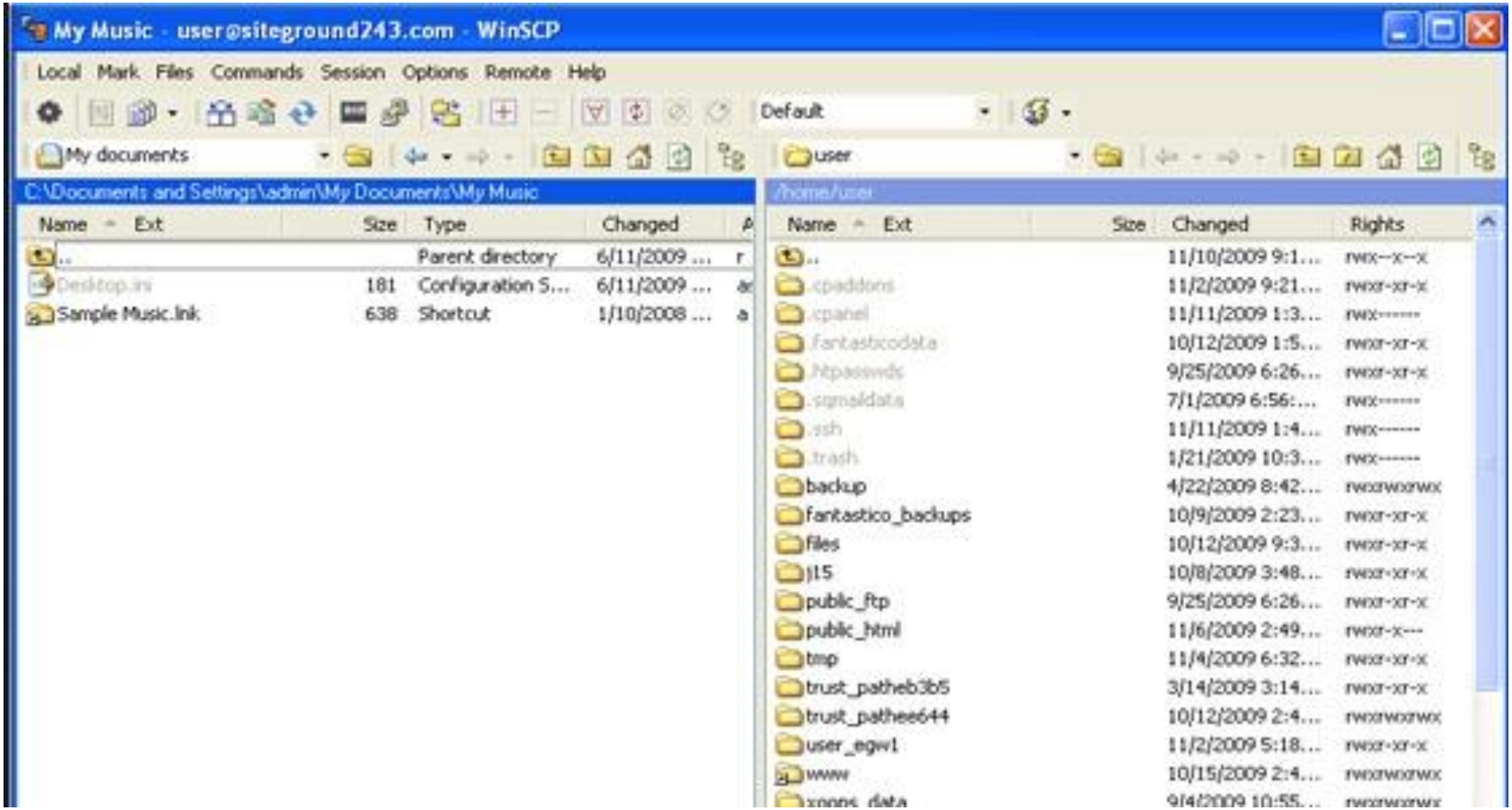
Remote server

montana.dataapplab.com

Data Application Lab




Data Application Lab



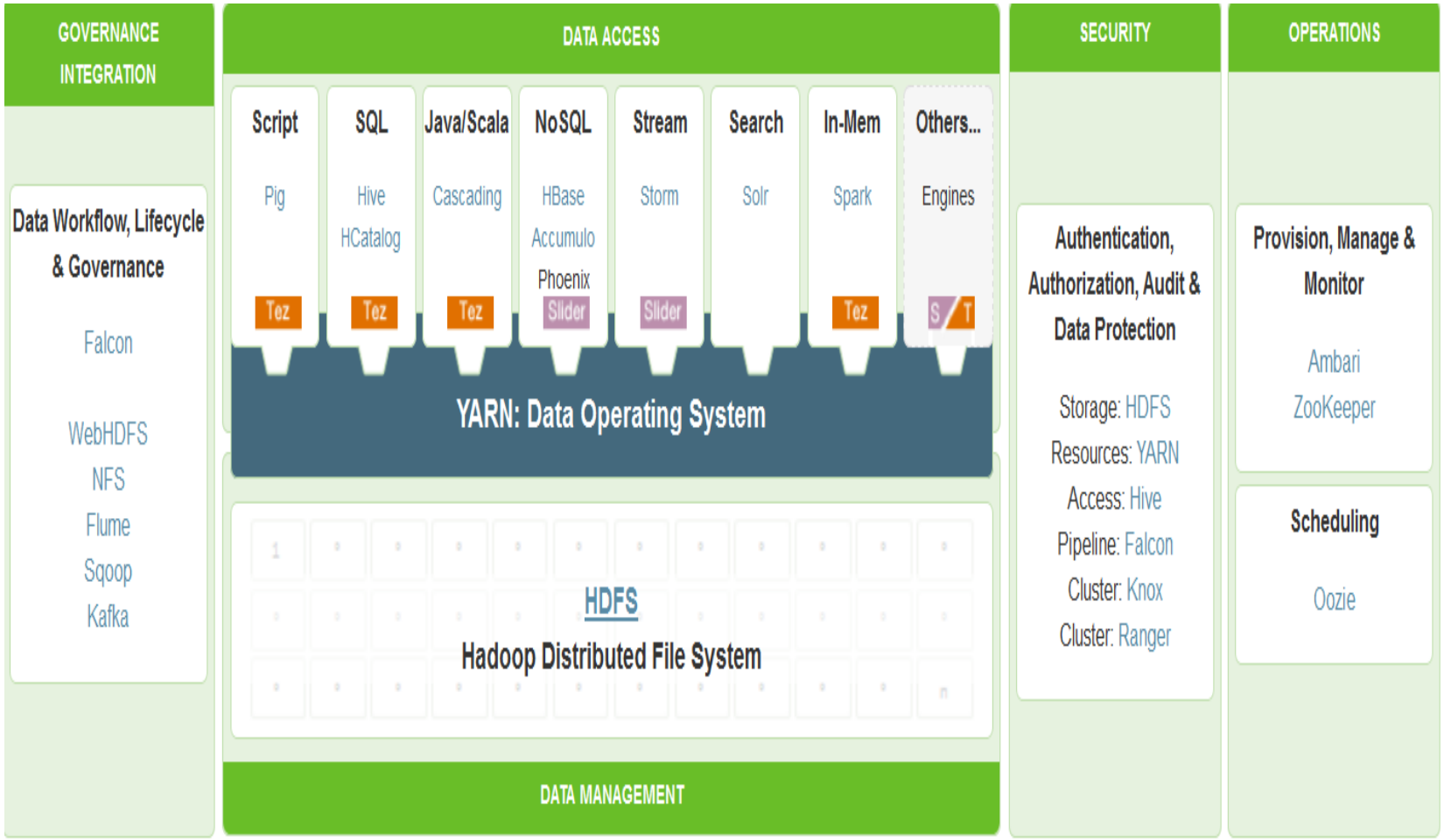

Data Application Lab

| | |
|---|---|
| **df [options]** | **Display used and available disk space.** |
| **du [options]** | Show how much space each file takes up. |
| **file [options] filename** | Determine what type of data is within a file. |
| **find [pathname] [expression]** | Search for files matching a provided pattern. |
| **grep [options] pattern [filesname]** | Search files or output for a particular pattern. |
| **kill [options] pid** | Stop a process. If the process refuses to stop, usekill -9 pid. |

| | |
|---|---|
| **less [options] [filename]** | **View the contents of a file one page at a time.** |
| **ln [options] source [destination]** | Create a shortcut. |
| **locate filename** | Search a copy of your filesystem for the specified filename. |
| **lpr [options]** | Send a print job. |
| **ls [options]** | List directory contents. |
| **man [command]** | Display the help information for the specified command. |
| **mkdir [options] directory** | Create a new directory. |
| **mv [options] source destination** | Rename or move file(s) or directories. |

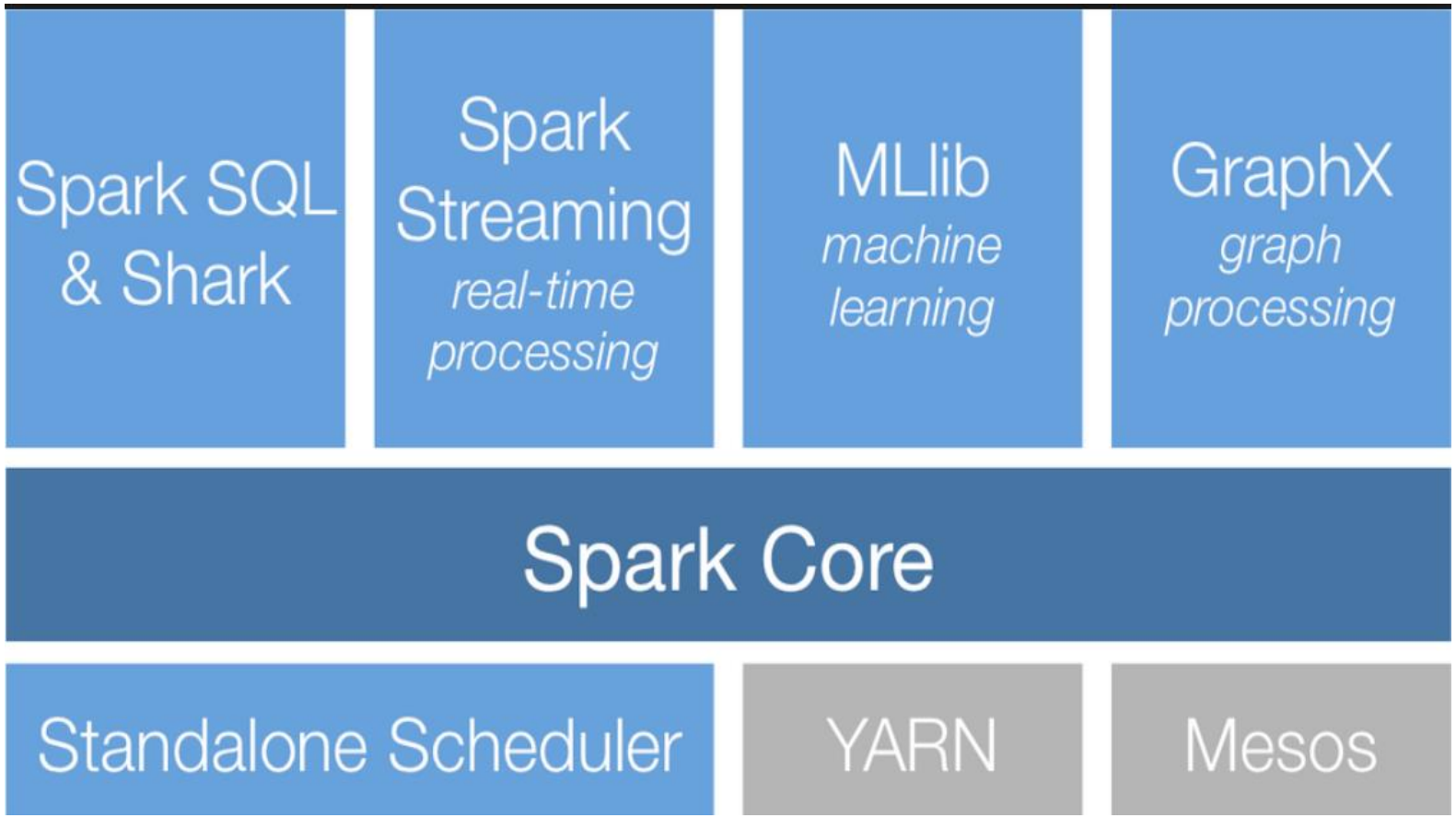
| passwd [name [password]] | Change the password or allow (for the system administrator) to change any password. |
|---|---|
| ps [options] | Display a snapshot of the currently running processes. |
| pwd | Display the pathname for the current directory. |
| rm [options] directory | Remove (delete) file(s) and/or directories. |
| rmdir [options] directory | Delete empty directories. |
| ssh [options] user@machine | Remotely log in to another Linux machine, over the network. Leave an ssh session by typing exit. |
| su [options] [user [arguments]] | Switch to another user account. |
| tail [options] [filename] | Display the last n lines of a file (the default is 10). |
| tar [options] filename | Store and extract files from a tarfile (.tar) or tarball (.tar.gz or .tgz). |
| top | Displays the resources being used on your system. Press q to exit. |
| touch filename | Create an empty file with the specified name. |
| who [options] | Display who is logged on. |

# Hortonworks



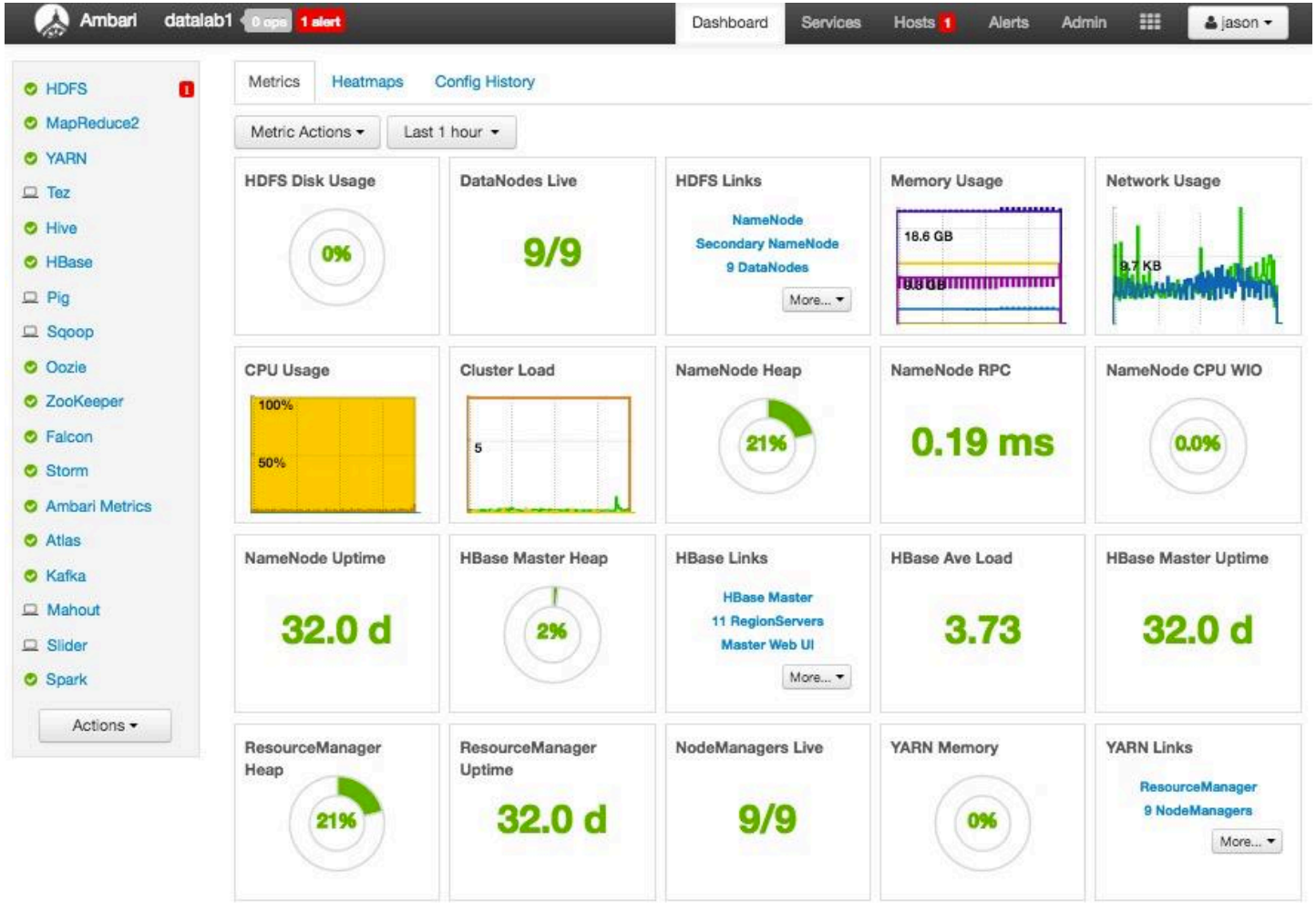

Data Application Lab

# Spark

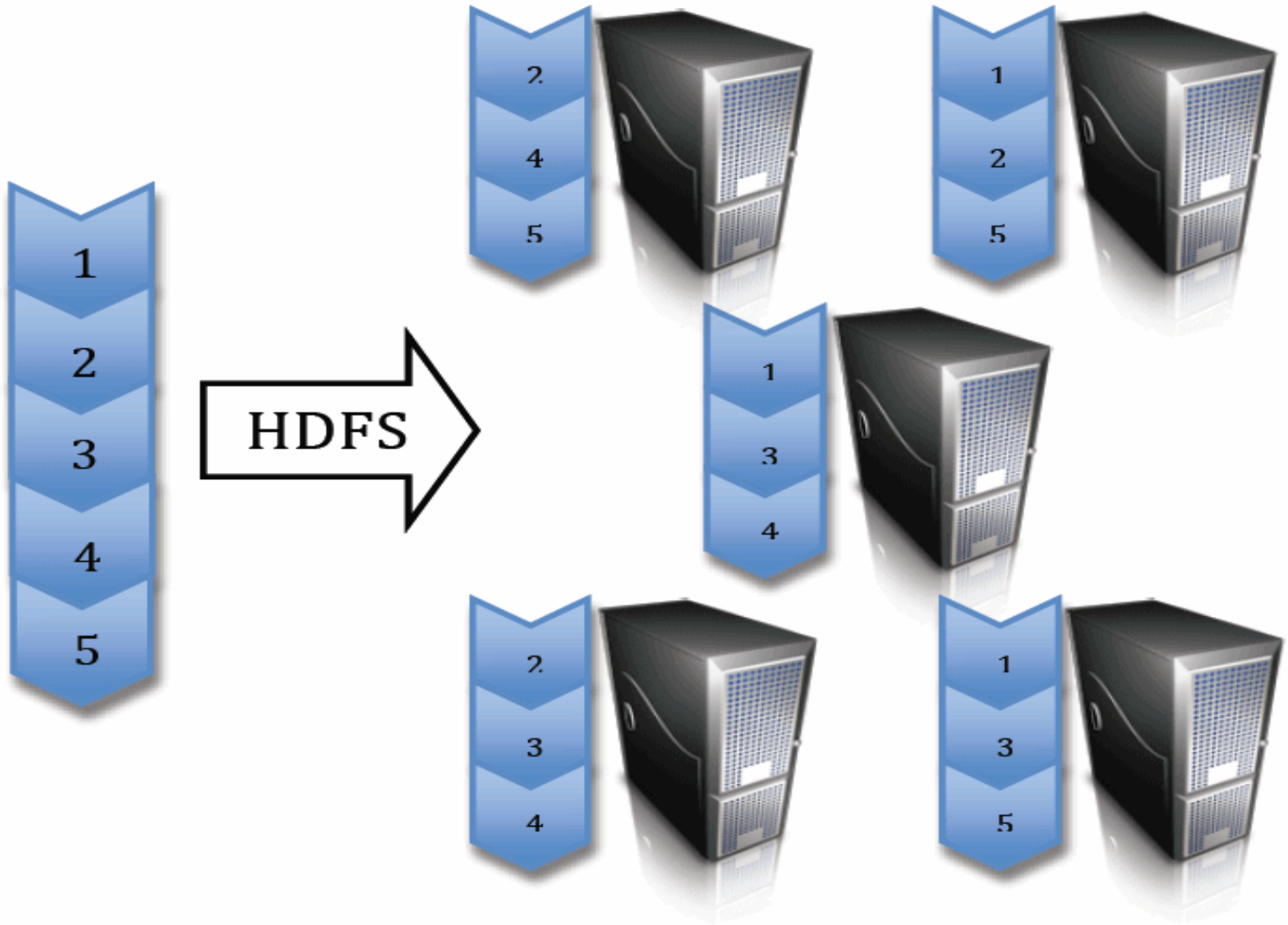



Data Application Lab

# DAL Cluster




Data Application Lab




Data Application Lab

# HDFS Commands

- hdfs dfs - ls
- Hdfs dfs – put
- Hdfs dfs – put /home/jason/test2.csv /user/jason/temp2

- Hdfs dfs – mkdir
- Hdfs dfs – rmdir
- Hdfs dfs – rm

- Reference :

https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

Data Application Lab

# Demo

# Thank you

Data Application Lab