

Thailand Big Data Challenge #2/2016

Apache Spark in Actions

18-19 June 2016

Dr.Thanachart Numnonda
IMC Institute
thanachart@imcinstitute.com

Outline

- Launch Azure Instance
- Install Docker on Ubuntu
- Pull Cloudera QuickStart to the docker
- HDFS
- Spark
- Spark SQL
- Spark Streaming

Cloudera VM

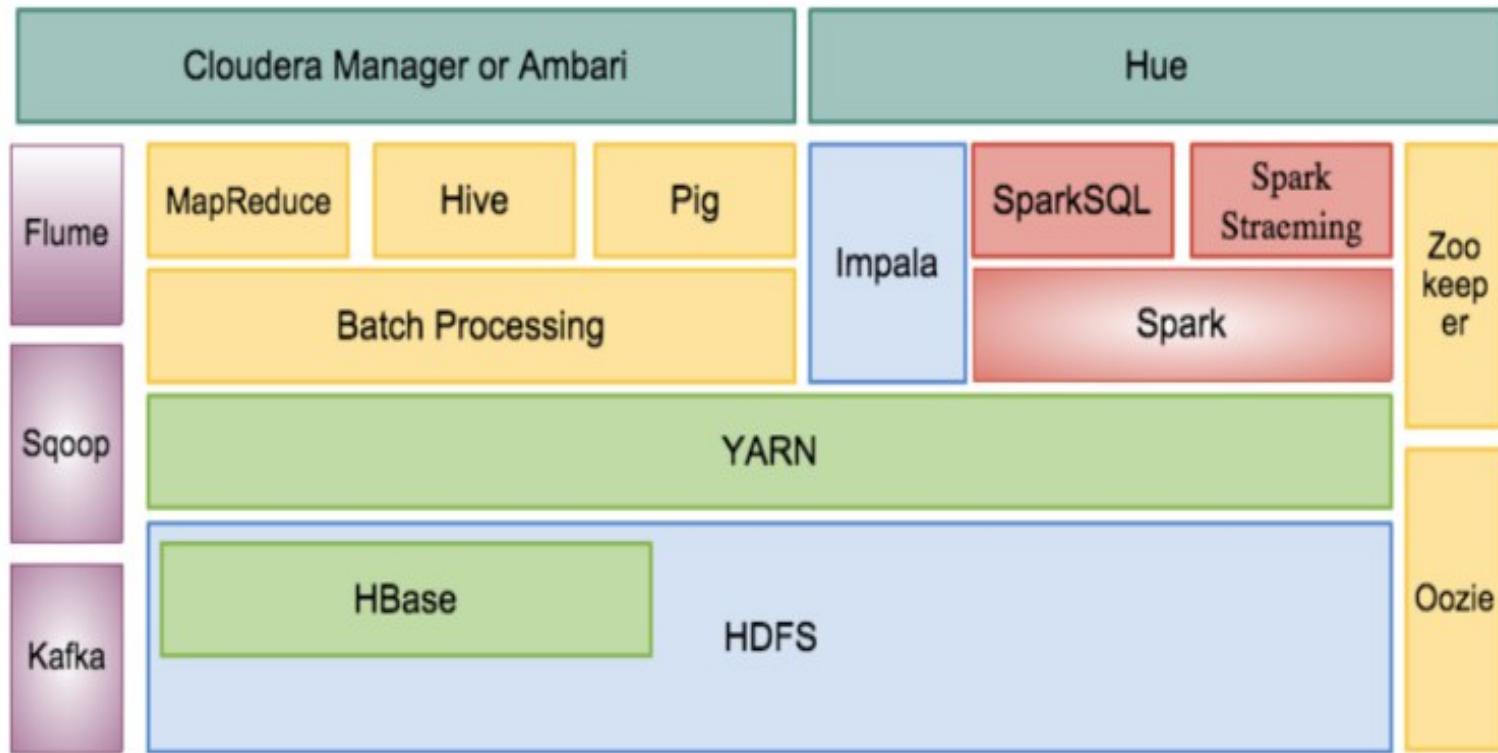
This lab will use a EC2 virtual server on AWS to install Cloudera. However, you can also use Cloudera QuickStart VM which can be downloaded from:

<http://www.cloudera.com/content/www/en-us/downloads.html>

The screenshot shows a landing page for downloading Cloudera software. The main title is "Download Cloudera Enterprise" with the subtitle "Local, On Premise, or Cloud-based Apache Hadoop Management". Below this, there are three large blue cards, each representing a different product:

- QuickStart VM**: Features an icon of two monitors. Description: "Get Started on your local machine using a QuickStart VM." Call-to-action buttons: "DOWNLOAD NOW" and "Learn More".
- Cloudera Manager**: Features an icon of a gear. Description: "A unified interface to manage your enterprise data hub. Express and Enterprise editions available." Call-to-action button: "DOWNLOAD NOW".
- Cloudera Director**: Features an icon of a cloud. Description: "Self-service, reliable experience for CDH and Cloudera Enterprise in the cloud" Call-to-action button: "DOWNLOAD NOW".

Hadoop Ecosystem



Spark Streaming

Hands-On: Launch a virtual server on Microsoft Azure

(Note: You can skip this session if you use your own computer or another cloud service)

Sign up for Visual Studio Dev Essential to get free Azure credit.

Showing: Visual Studio Dev Essentials

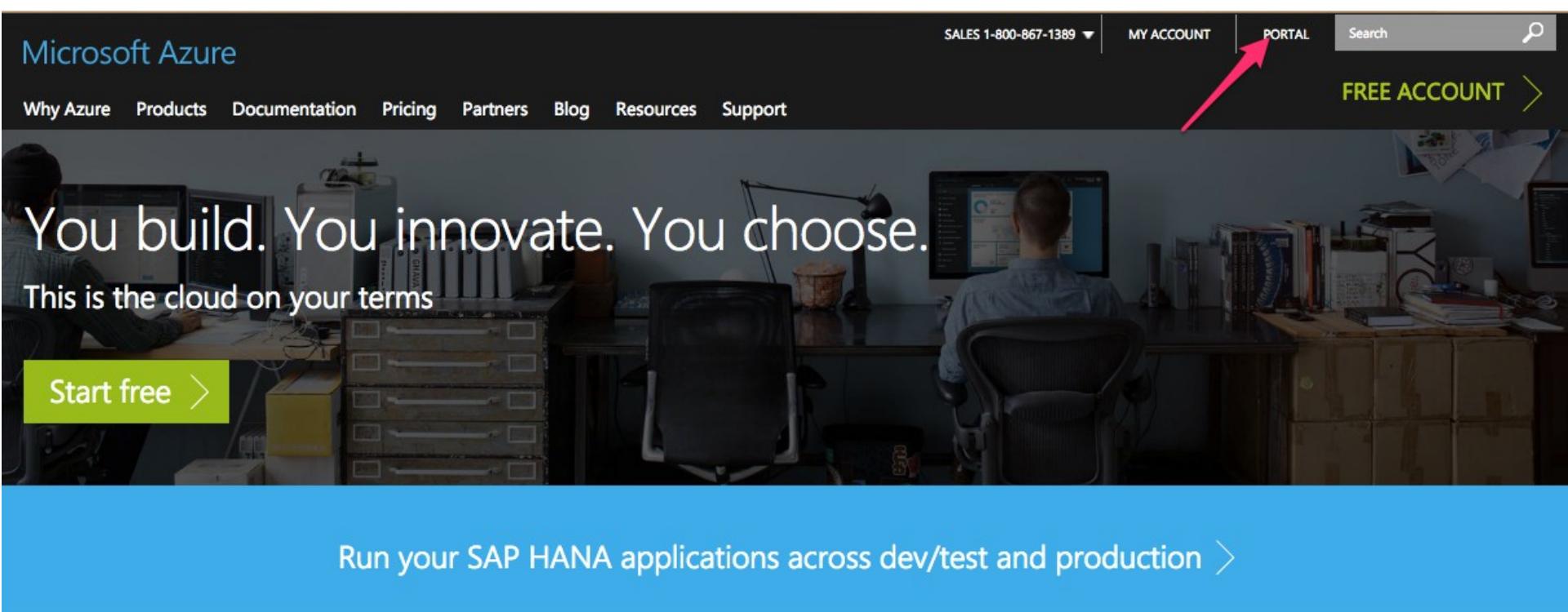
Visual Studio Dev Essentials



Featured (6)

| | | | | | |
|---|---|---|---|--|---|
|  Visual Studio Community Full-featured, extensible IDE Free for individuals, open source or small teams. Create apps for Windows, iOS, Andro... See more |  Visual Studio Code Modern lightweight editor A powerful, streamlined code editor for your favorite platform - Linux, Mac OS X, and... Windows |  Visual Studio Team Services Basic level Free Git repos, Agile planning tools and hosted builds, for any language – it's the perfect... complement to your IDE |  Azure \$25 monthly credit for 1 year Your own personal sandbox for dev/test! VMs, cloud services, and more. Credit cannot be... applied to existing Azure |  Xamarin University Training Free on-demand access Build native iOS and Android apps in C# with expert getting-started videos (subset of class... videos and materials) |  Pluralsight 3-month subscription World-class training taught by an elite group of industry leaders. |
| Download  | Download  | Get started  | Activate  | Get Code  | Get Code  |

Sign in to Azure Portal



The screenshot shows the Microsoft Azure homepage. At the top right, there is a navigation bar with links for "SALES 1-800-867-1389", "MY ACCOUNT", "PORTAL" (which has a red arrow pointing to it), and a search bar. Below the navigation bar, there is a banner with the text "You build. You innovate. You choose." and "This is the cloud on your terms". A green button labeled "Start free >" is visible. In the background, there is a photograph of a person working at a desk with multiple monitors. At the bottom of the page, there is a blue bar with the text "Run your SAP HANA applications across dev/test and production >".

Microsoft Azure ▾

Search resources

1

New dashboard Edit dashboard Share Fullscreen Clone Delete

Dashboard

All resources ALL SUBSCRIPTIONS

- hdp41n71a7p
- imclabstorage
- egahdpstorage

Service health MY RESOURCES



Marketplace

Subscriptions Forecast expenses and costs to optimize your apps

Help + support

Feedback

Information icons

The Microsoft Azure dashboard interface. At the top, there's a search bar labeled "Search resources" and various navigation icons. Below the header, the title "Dashboard" is displayed with options to "New dashboard", "Edit dashboard", "Share", "Fullscreen", "Clone", and "Delete". A sidebar on the left lists "All resources" under "ALL SUBSCRIPTIONS" with three items: "hdp41n71a7p", "imclabstorage", and "egahdpstorage". The main area features a "Service health" section with a world map showing green checkmarks across most regions. Below this are four large cards: "Marketplace" (blue shopping bag icon), "Subscriptions" (green and yellow dollar sign icon with text "Forecast expenses and costs to optimize your apps"), "Help + support" (blue lock icon), and "Feedback" (red heart icon). At the bottom, there are three blue circular icons with white "i" symbols and a blue gear icon.

Virtual Server

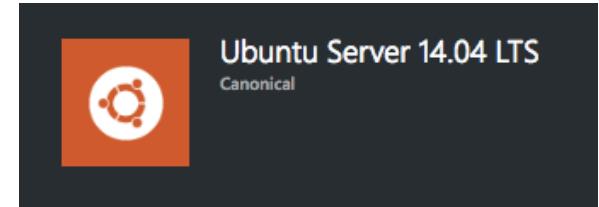
This lab will use an Azure virtual server to install a Cloudera Quickstart Docker using the following features:

Ubuntu Server 14.04 LTS

DS3_V2 Standard 4 Core, 14 GB memory, 28 GB SSD

Select New => Virtual Machines => Virtual Machines

The screenshot shows the Microsoft Azure portal interface. On the left, there's a sidebar with various icons and a search bar labeled "Search the marketplace". Below the search bar, under "MARKETPLACE", the "Virtual Machines" option is highlighted with a pink arrow. The main area is titled "Virtual Machines" and displays a list of "FEATURED APPS". The first item is "Windows Server 2012 R2 Datacenter". The second item, "Ubuntu Server 14.04 LTS", is also highlighted with a pink arrow. Below it are "SQL Server 2014 Enterprise on Windows Server 2012 R2" and "SharePoint 2013 HA Farm".



Ubuntu Server 14.04.4 LTS (amd64 20160516) for Microsoft popular Linux for cloud environments. Updates and patches until 2019-04-17. Ubuntu Server is the perfect virtual machine for web applications to NoSQL databases and Hadoop. For more information, visit [the documentation](#) or [using Juju to deploy your workloads](#).

Legal Terms

By clicking the Create button, I acknowledge that I am getting the [legal terms](#) of Canonical apply to it. Microsoft does not accept responsibility for the content of these pages. Also see the [privacy statement](#) from Canonical.

This screenshot shows the "Create" page for the Ubuntu Server 14.04 LTS deployment. At the top, there's a row of social media sharing icons. Below them is a dropdown menu labeled "Select a deployment model" with "Resource Manager" selected. A large red arrow points to the prominent blue "Create" button at the bottom of the page.

On the Basics page, enter:

- a name for the VM
- a username for the Admin User
- the Authentication Type set to password
- a password
- a resource group name

Microsoft Azure New > Virtual Machines > Ubuntu Server 14.04 LTS > Create virtual machine > Basics

+ New

Resource groups

All resources

Recent

App Services

SQL databases

Virtual machines (classic)

Virtual machines

Cloud services (classic)

Subscriptions

Storage accounts (classic...)

Browse >

Create virtual machine

1 Basics Configure basic settings

2 Size Choose virtual machine size

3 Settings Configure optional features

4 Summary Ubuntu Server 14.04 LTS

Basics

* Name clouderadocker ✓

* User name imcinstiute ✓

* Authentication type Password SSH public key

* Password ✓

Subscription Developer Program Benefit

* Resource group ⓘ

Create new Use existing

OK

Choose DS3_v2 Standard

Microsoft Azure < Virtual Machines > Ubuntu Server 14.04 LTS > Create virtual machine > Choose a size

Create virtual machine

Choose a size

Browse the available sizes and their features

| DS1_V2 Standard | DS2_V2 Standard | DS3_V2 Standard |
|---------------------------------------|--|--|
| 1 Core | 2 Cores | 4 Cores → |
| 3.5 GB | 7 GB | 14 GB |
| 2 Data disks | 4 Data disks | 8 Data disks |
| 3200 Max IOPS | 6400 Max IOPS | 12800 Max IOPS |
| 7 GB Local SSD | 14 GB Local SSD | 28 GB Local SSD |
| Load balancing | Load balancing | Load balancing |
| Auto scale | Auto scale | Auto scale |
| Premium disk support | Premium disk support | Premium disk support |
| 67.70 USD/MONTH (ESTIMATED) | 128.71 USD/MONTH (ESTIMATED) | 257.42 USD/MONTH (ESTIMATED) |

1 Basics Done

2 Size Choose virtual machine size >

3 Settings Configure optional features >

4 Summary Ubuntu Server 14.04 LTS >

Select

The screenshot shows the Microsoft Azure 'Create virtual machine' wizard at the 'Choose a size' step. On the left, a sidebar lists various services: App Service, Functions, Logic Apps, Container Registry, Storage, Database, and more. The main area shows a progress bar with four steps: 1. Basics (Done), 2. Size (selected), 3. Settings, and 4. Summary. The 'Size' step is titled 'Choose virtual machine size'. To the right, a table compares three VM sizes: DS1_V2 Standard, DS2_V2 Standard, and DS3_V2 Standard. The DS3_V2 Standard row is highlighted with a blue background. A red arrow points to the 'Cores' column for DS3_V2 Standard, which is listed as '4'. Other columns show memory (14 GB vs 7 GB vs 3.5 GB), data disks (8 vs 4 vs 2), and IOPS (12800 vs 6400 vs 3200). Estimated monthly costs are listed at the bottom of each row: \$67.70 for DS1, \$128.71 for DS2, and \$257.42 for DS3.

Microsoft Azure New > Virtual Machines > Ubuntu Server 14.04 LTS > Create virtual machine

The screenshot shows the Microsoft Azure 'Create virtual machine' wizard. The left sidebar lists various service icons. The main window displays four steps:

- 1 Basics**: Done.
- 2 Size**: Done.
- 3 Settings**: Configure optional features. This step is highlighted with a light blue background.
- 4 Summary**: Ubuntu Server 14.04 LTS.

The 'Settings' tab is open, showing storage and network configurations. The 'Storage' section includes:

- Disk type: Premium (SSD) (selected, highlighted with a blue border and a red arrow pointing to it).
- * Storage account: (new) defaultstoragesouthc6264.

The 'Network' section includes:

- * Virtual network: (new) Default-Storage-SouthCentralUS.
- * Subnet: default (10.0.0.0/24).

An 'OK' button is at the bottom right of the settings tab.

Microsoft Azure New > Virtual Machines > Ubuntu Server 14.04 LTS > Create virtual machine > Summary

Create virtual machine

Summary

1 Basics Done ✓

2 Size Done ✓

3 Settings Done ✓

4 Summary Ubuntu Server 14.04 LTS >

Validation passed

Basics

| | |
|----------------|--|
| Subscription | Developer Program Benefit |
| Resource group | Default-MachineLearning-SouthCentralUS |
| Location | South Central US |

Settings

| | |
|-----------------|--|
| Computer name | clouderadocker |
| User name | imcinstitute |
| Size | Standard DS11 v2 |
| Disk type | Premium (SSD) |
| Storage account | (new) defaultmachinelearn2038 |
| Virtual network | (new) Default-MachineLearning-SouthCentralUS |
| Subnet | (new) default (10.1.0.0/24) |

OK



Microsoft Azure cluderadocker > Settings

cluderadocker
Virtual machine

Settings Connect Start Restart Stop Delete

Essentials

Resource group
[Default-MachineLearning-SouthCentralU...](#)

Status
Running

Location
South Central US

Subscription name
[Developer Program Benefit](#)

Subscription ID
59cbb519-5261-48e5-9825-9df96f8302a9

Computer name
cluderadocker

Operating system
Linux

Size
Standard DS11 v2 (2 cores, 14 GB memory)

Public IP address/DNS name label
[104.210.146.182/<none>](#)

Virtual network/subnet
[Default-MachineLearning-SouthCentralUS/...](#)

All settings →

Monitoring

CPU percentage

100%

Add tiles +

Settings

cluderadocker

Filter settings

SUPPORT + TROUBLESHOOTING

- [Troubleshoot](#)
- [Audit logs](#)
- [Resource health](#)
- [Boot diagnostics](#)
- [Reset password](#)
- [Redeploy](#)
- [New support request](#)

GENERAL

Setting the inbound port for Hue (8888)

The screenshot shows two windows side-by-side in the Azure portal.

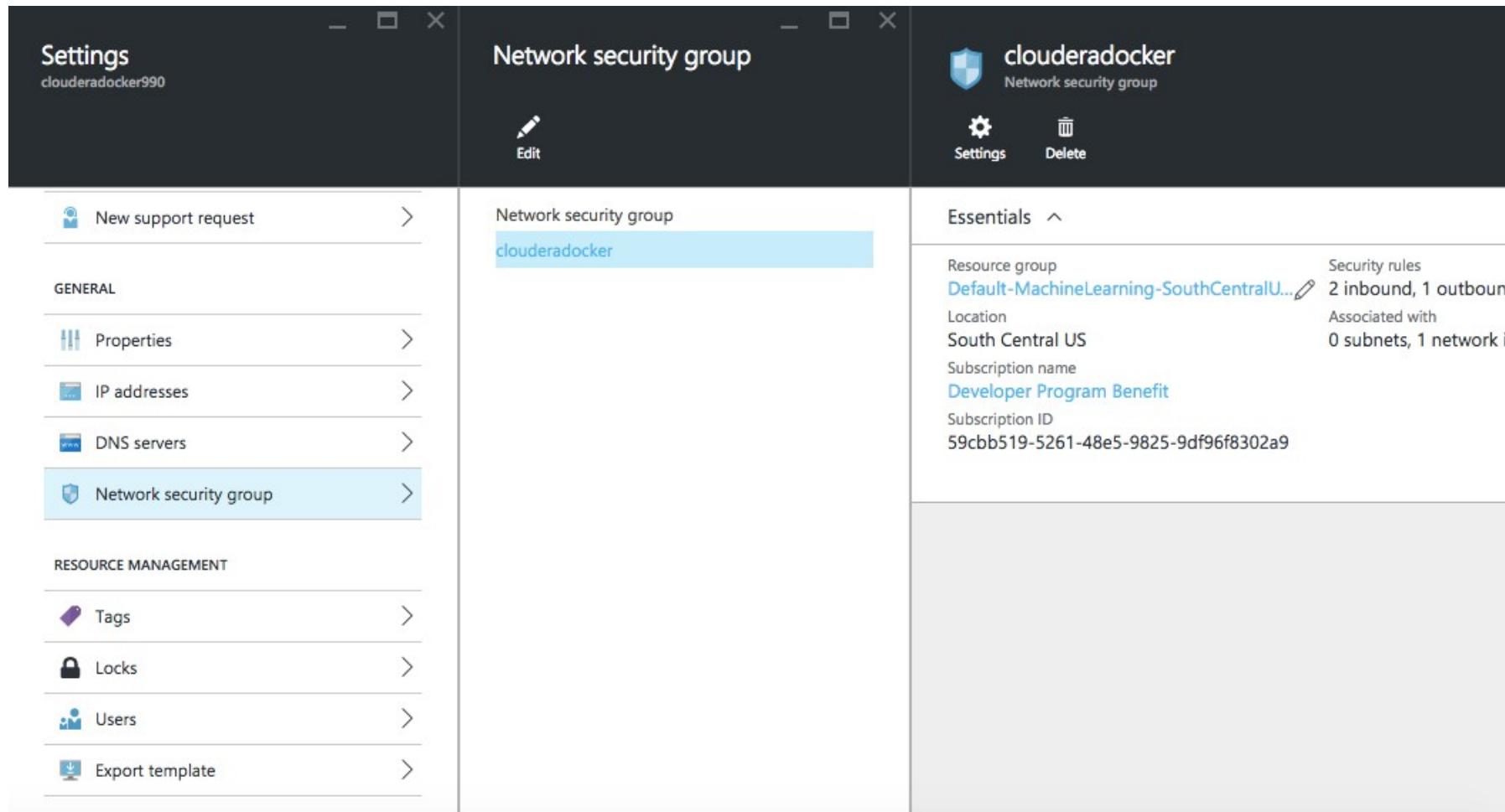
Left Window: Settings

- General:
 - Properties
 - Disks
 - Network interfaces** (highlighted in blue)
 - Availability set
 - Extensions
 - Size
- Monitoring:
 - Alert rules
 - Diagnostics
- Resource Management

Right Window: Network interfaces

A table displays network interface information:

| NAME | PUBLIC IP ADDRE... | PRIVATE IP ADDR... | SECURITY GROUP | ... |
|-------------------|--------------------|--------------------|----------------|-----|
| clouderadocker990 | 104.210.146.182 | 10.1.0.4 | clouderadocker | ... |



The screenshot shows three separate windows from the Microsoft Azure portal:

- Settings** window (left): Shows a list of options under "GENERAL" and "RESOURCE MANAGEMENT". The "Network security group" option is selected.
- Network security group** window (middle): Displays a list of network security groups, with "clouderadocker" highlighted.
- clouderadocker** window (right): Provides details for the selected network security group, including its resource group, location, subscription information, and security rules.

Settings

clouderadocker

SUPPORT + TROUBLESHOOTING

- Audit logs >
- New support request >

GENERAL

- Properties >
- Inbound security rules > hue
- Outbound security rules >
- Network interfaces >
- Subnets >

Inbound security rules

clouderadocker

+
Add
Default rules

■ Search inbound security rules

| PRIORITY | NAME | SOURCE | DESTINATION | SERVICE |
|----------|-------------------|--------|--|--|
| 1000 | default-allow-ssh | Any | ■ hue Default-MachineLearning-SouthCentralUS | ■ Save ■ Discard ■ Delete |

* Name

* Priority

* Source Any CIDR block Tag

* Protocol Any TCP UDP

* Source port range

* Destination Any CIDR block Tag

* Destination port range

* Action

DEPLOY DIRECTORY

Inbound security rules

clouderadocker

Add **Default rules**

Filter settings

SUPPORT + TROUBLESHOOTING

- Audit logs >
- New support request >

GENERAL

- Properties >
- Inbound security rules > **Selected**
- Outbound security rules >
- Network interfaces >
- Subnets >

Search inbound security rules

| PRIORITY | NAME | SOURCE | DESTINATION | SERVICE |
|----------|-------------------|--------|-------------|--------------|
| 1000 | default-allow-ssh | Any | Any | SSH (TCP/22) |

Get the IP address

The screenshot displays two side-by-side views in the Azure portal:

Left View (Virtual Machine):

- Resource group:** Default-MachineLearning-SouthCentralU...
- Status:** Running
- Location:** South Central US
- Subscription name:** Developer Program Benefit
- Subscription ID:** 59ccb519-5261-48e5-9825-9df96f8302a9
- Computer name:** cluderadocker
- Operating system:** Linux
- Size:** Standard DS11 v2 (2 cores, 14 GB memory)
- Public IP address/DNS name label:** 104.210.146.182/<none>
- Virtual network/subnet:** Default-MachineLearning-SouthCentralUS/...

Right View (Public IP address):

- Resource group:** Default-MachineLearning-SouthCentralU...
- Location:** South Central US
- Subscription name:** Developer Program Benefit
- Subscription ID:** 59ccb519-5261-48e5-9825-9df96f8302a9
- IP address:** 104.210.146.182
- DNS name:** -
- Associated to:** cluderadocker990
- Virtual machine:** cluderadocker

Connect to an instance from Mac/Linux

```
ssh -i ~/.ssh/id_rsa imcinstigate@104.210.146.182
```

WARNING! Your environment specifies an invalid locale.

This can affect your user experience significantly, including the ability to manage packages. You may install the locales by running:

```
sudo apt-get install language-pack-UTF-8  
or  
sudo locale-gen UTF-8
```

To see all available language packs, run:

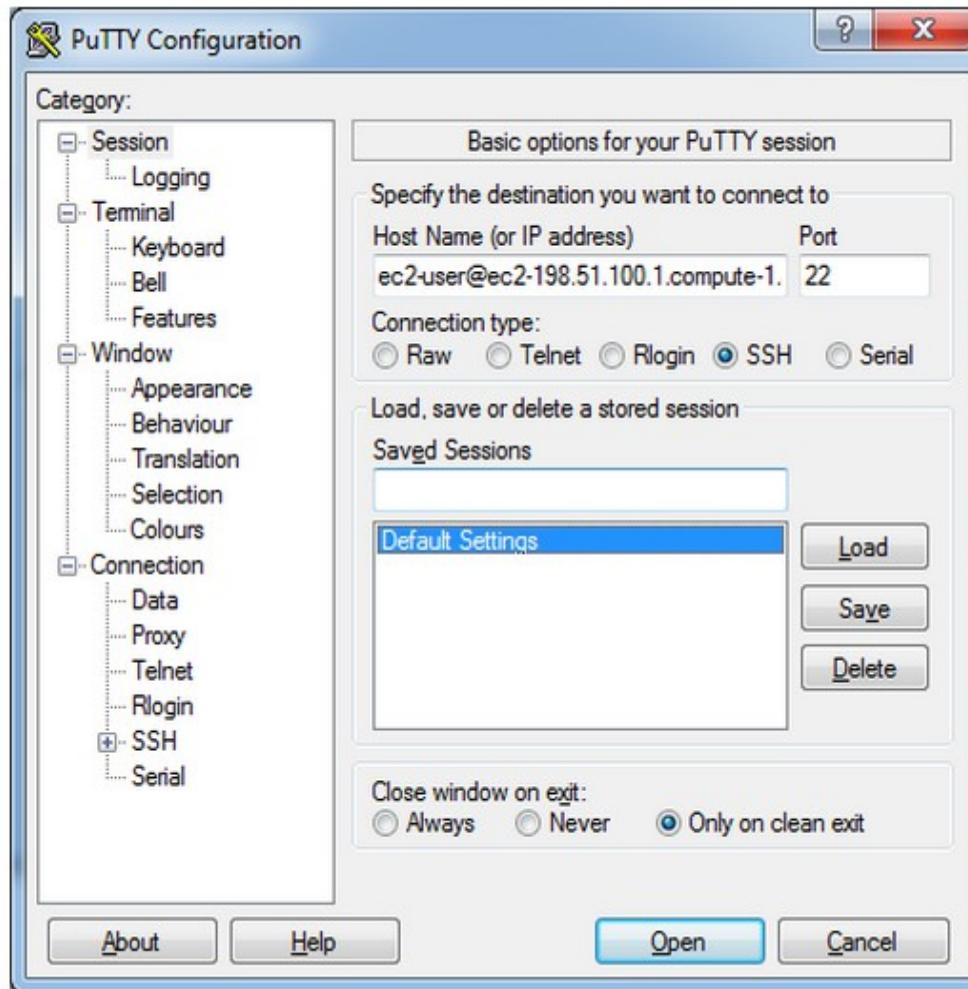
```
apt-cache search "^language-pack-[a-z][a-z]$"
```

To disable this message for all users, run:

```
sudo touch /var/lib/cloud/instance/locale-check.skip
```

```
imcinstigate@clouderadocker:~$ █
```

Connect to an instance from Windows using Putty



Hands-On: Installing Cloudera Quickstart on Docker Container

Installation Steps

- Update OS
- Install Docker
- Pull Cloudera Quickstart
- Run Cloudera Quickstart
- Run Cloudera Manager

Update OS (Ubuntu)

- Command: sudo apt-get update

```
ubuntu@ip-172-31-30-238:~$ sudo apt-get update
Ign http://us-east-1.ec2.archive.ubuntu.com trusty InRelease
Get:1 http://us-east-1.ec2.archive.ubuntu.com trusty-updates InRelease [65.9 kB]
Get:2 http://us-east-1.ec2.archive.ubuntu.com trusty-backports InRelease [65.9 kB]
Hit http://us-east-1.ec2.archive.ubuntu.com trusty Release.gpg
Hit http://us-east-1.ec2.archive.ubuntu.com trusty Release
Get:3 http://security.ubuntu.com trusty-security InRelease [65.9 kB]
Get:4 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/main Sources [277 kB]
Get:5 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/restricted Sources [535
2 B]
Get:6 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/universe Sources [156 k
B]
Get:7 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/multiverse Sources [593
9 B]
Get:8 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/main amd64 Packages [78
1 kB]
```

Docker Installation

- Command: sudo apt-get install docker.io

```
ubuntu@ip-172-31-30-238:~$ sudo apt-get install docker.io
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  aufs-tools cgroup-lite git git-man liberror-perl
Suggested packages:
  btrfs-tools debootstrap lxc rinse git-daemon-run git-daemon-sysvinit git-doc
  git-el git-email git-gui gitk gitweb git-arch git-bzr git-cvs git-mediawiki
  git-svn
The following NEW packages will be installed:
  aufs-tools cgroup-lite docker.io git git-man liberror-perl
0 upgraded, 6 newly installed, 0 to remove and 84 not upgraded.
Need to get 8150 kB of archives.
After this operation, 51.4 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu/ trusty/universe aufs-tools amd
64 1:3.2+20130722-1.1 [92.3 kB]
```

Pull Cloudera Quickstart

- Command: sudo docker pull cloudera/quickstart:latest

```
ubuntu@ip-172-31-30-238:~$ sudo docker pull cloudera/quickstart:latest
latest: Pulling from cloudera/quickstart
2cda82941cb7: Already exists
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Downloaded newer image for cloudera/quickstart:latest
ubuntu@ip-172-31-30-238:~$ █
```

Show docker images

- Command: sudo docker images

```
ubuntu@ip-172-31-30-238:~$ sudo docker images
REPOSITORY          TAG      IMAGE ID      CREATED
VIRTUAL SIZE
cloudera/quickstart    latest  2cda82941cb7  9 weeks ago
6.336 GB
```

Run Cloudera quickstart

- Command: `sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i [OPTIONS] [IMAGE] /usr/bin/docker-quickstart`

Example: `sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888:8888 cloudera/quickstart /usr/bin/docker-quickstart`

```
ubuntu@ip-172-31-30-238:~$ sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888:8888 -p 7180:7180 cloudera/quickstart /usr/bin/docker-quickstart
Starting mysqld: [ OK ]  
  
if [ "$1" == "start" ] ; then
  if [ "${EC2}" == 'true' ] ; then
    FIRST_BOOT_FLAG=/var/lib/cloudera-quickstart/.ec2-key-installed
    if [ ! -f "${FIRST_BOOT_FLAG}" ] ; then
      METADATA_API=http://169.254.169.254/latest/meta-data
      KEY_URL=${METADATA API}/public-keys/0/openssh-key
```

Docker commands:

- *docker images*
- *docker ps*
- *docker attach id*
- *docker kill id*
- *Exit from container*
 - *exit (exit & kill the running image)*
 - *Ctrl-P, Ctrl-Q (exit without killing the running image)*

Login to Hue

http://104.210.146.182:8888

Welcome to Hue
Sign in to continue to your dashboard



Username

Password

Sign in

Hue and the Hue logo are trademarks of Cloudera, Inc.

Quick Start Wizard - Hue™ 3.9.0 - The Hadoop UI

Step 1: Check Configuration

Step 2: Examples

Step 3: Users

Step 4: Go!

Checking current configuration

Configuration files located in [`/etc/hue/conf.empty`](#)

All OK. Configuration check passed.

Back

Next

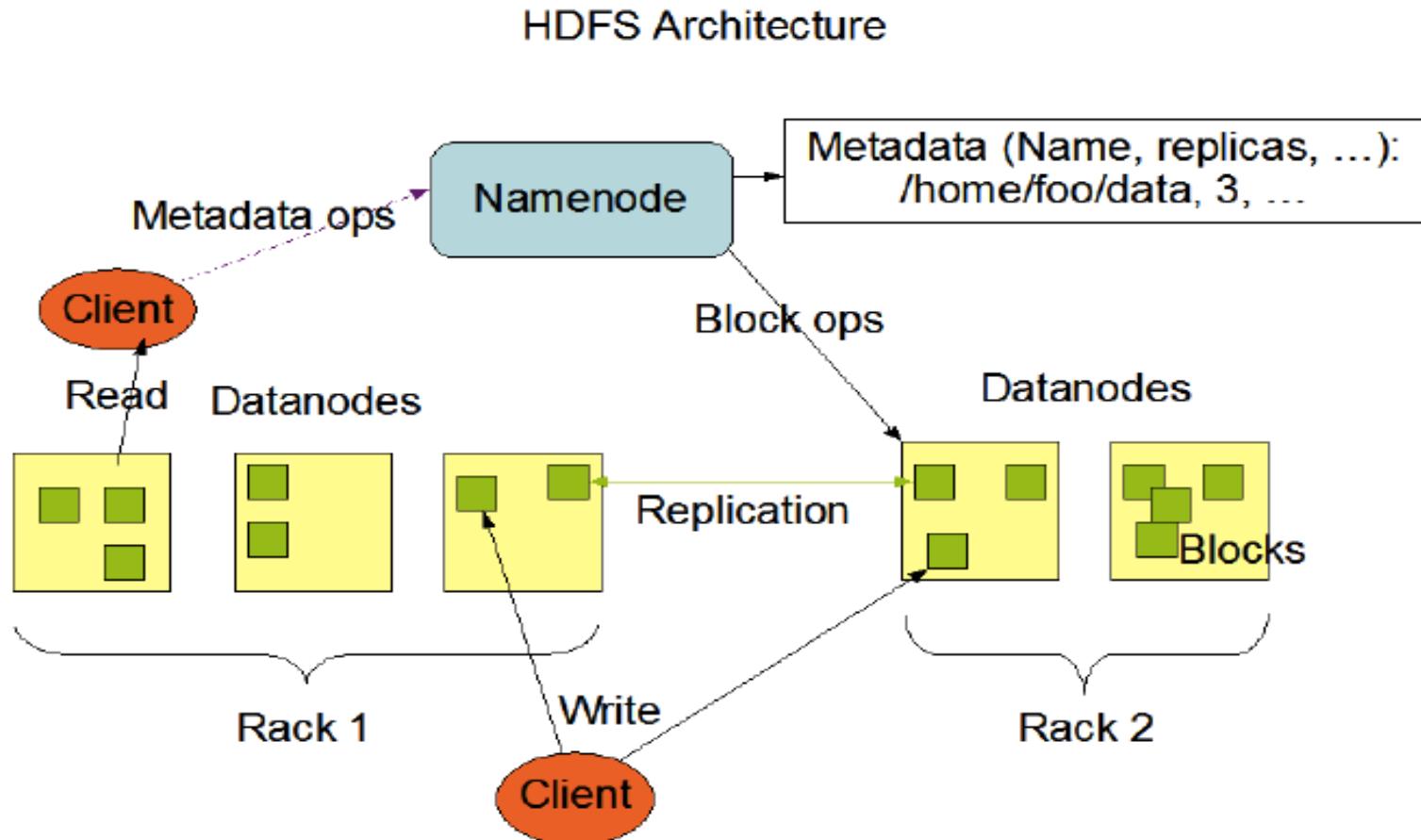
Hue and the Hue logo are trademarks of Cloudera, Inc.

Hands-On: Importing/Exporting Data to HDFS

HDFS

- Default storage for the Hadoop cluster
- Data is distributed and replicated over multiple machines
- Designed to handle very large files with streaming data access patterns.
- NameNode/DataNode
- Master/slave architecture (1 master 'n' slaves)
- Designed for large files (64 MB default, but configurable) across all the nodes

HDFS Architecture



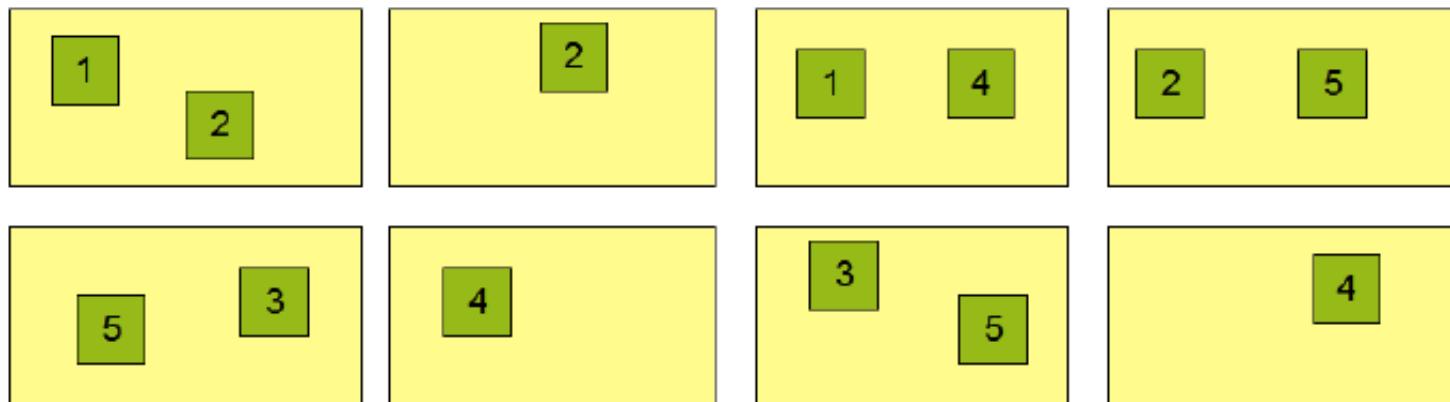
Source Hadoop: Shashwat Shriparv

Data Replication in HDFS

Block Replication

```
Namenode (Filename, numReplicas, block-ids, ...)  
/users/sameerp/data/part-0, r:2, {1,3}, ...  
/users/sameerp/data/part-1, r:3, {2,4,5}, ...
```

Datanodes



Source Hadoop: Shashwat Shriparv

How does HDFS work?

A file we want to store on HDFS ...

600 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

We've read over and over again about Nash refusing to ask for a trade, refusing to play the game that so many others have late in their careers.

Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?

HDFS Splits file into **blocks** ...

256 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

256 MB

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

88 MB

We've read over and over again about Nash refusing to play the game that so many others have late in their careers.

Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?

HDFS will create **3replicas** of each block ...

3 copies

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

3 copies

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

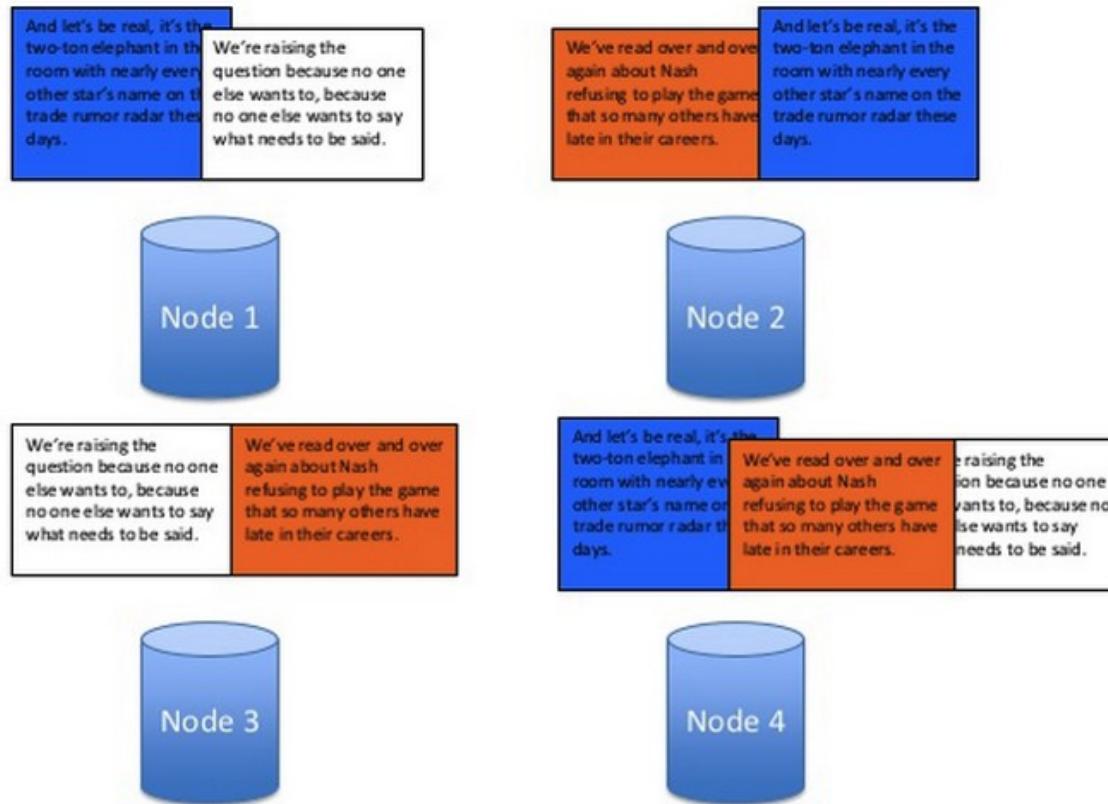
3 copies

We've read over and over again about Nash refusing to play the game that so many others have late in their careers.

Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?

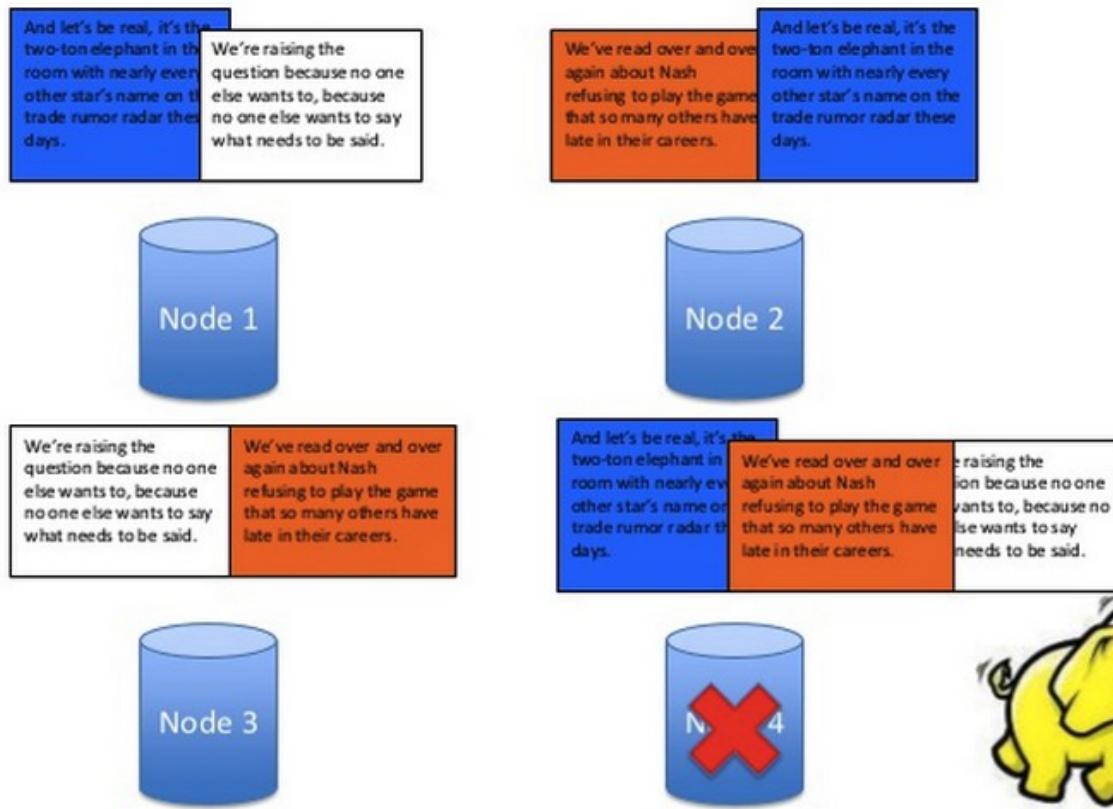
HDFS distributes these replicas across the cluster ...



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

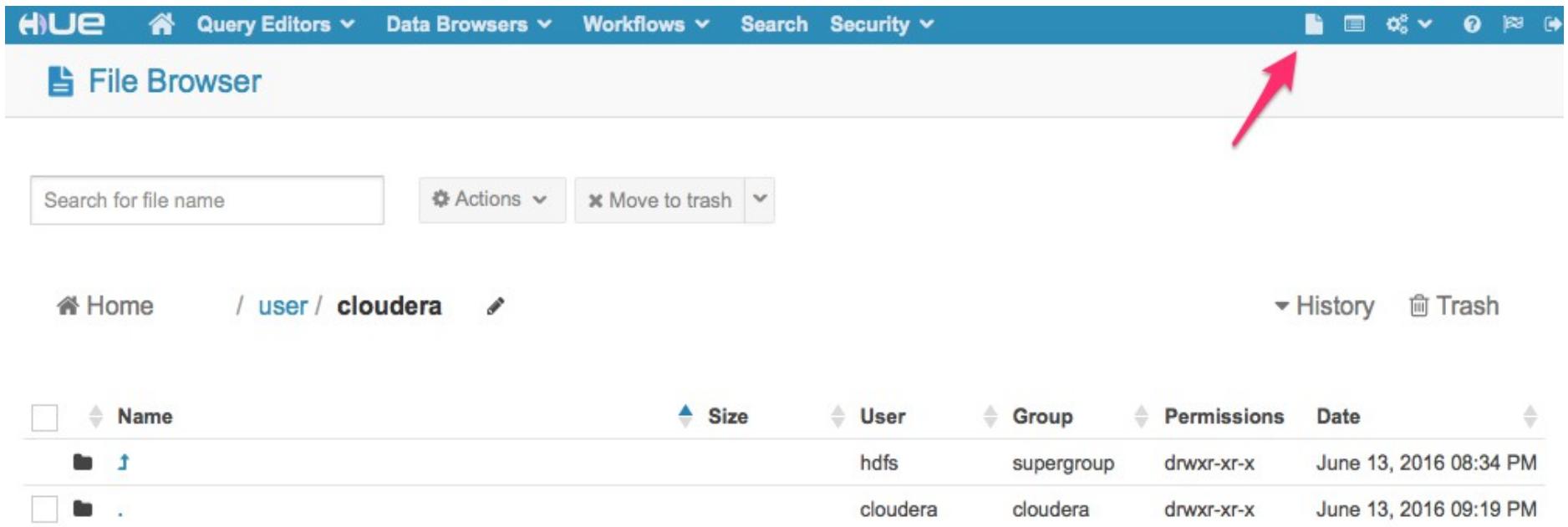
How does HDFS work?

If a node goes down, we have copies elsewhere



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

Review file in Hadoop HDFS using File Browse



The screenshot shows the Hue File Browser interface. At the top, there is a navigation bar with links for Home, Query Editors, Data Browsers, Workflows, Search, Security, and various system icons. Below the navigation bar, the title "File Browser" is displayed next to a folder icon. On the left, there is a search bar labeled "Search for file name" and a "Actions" dropdown menu. In the center, the current path is shown as "/ user / cloudera". To the right of the path are "History" and "Trash" buttons. The main area displays a table of files in the "/user/cloudera" directory. The table has columns for Name, Size, User, Group, Permissions, and Date. Two files are listed:

| Name | Size | User | Group | Permissions | Date |
|------|------|----------|------------|-------------|------------------------|
| .. | | hdfs | supergroup | drwxr-xr-x | June 13, 2016 08:34 PM |
| . | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:19 PM |

Create a new directory name as: **input & output**

The screenshot shows the Hue File Browser interface. At the top, there is a navigation bar with links for Home, Query Editors, Data Browsers, Workflows, Search, and Security. Below the navigation bar, the title "File Browser" is displayed. On the left, there is a sidebar with "Actions" and "Move to trash" dropdown menus, and "Upload" and "New" buttons. The main area shows a list of files and directories under the path "/user/cloudera". The list includes two entries: "hdfs" (User: supergroup, Group: supergroup, Permissions: drwxr-xr-x, Date: June 13, 2016 08:34 PM) and "cloudera" (User: cloudera, Group: cloudera, Permissions: drwxr-xr-x, Date: June 13, 2016 09:19 PM). To the right of the list, there are "History" and "Trash" buttons. A red arrow points to the "New" button, which has a dropdown menu open. The dropdown menu contains two options: "File" and "Directory". The "Directory" option is highlighted with a red box and a red arrow pointing to it.

| Size | User | Group | Permissions | Date |
|------|----------|------------|-------------|------------------------|
| | hdfs | supergroup | drwxr-xr-x | June 13, 2016 08:34 PM |
| | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:19 PM |

Directory Name

Create

HUE  Query Editors ▾ Data Browsers ▾ Workflows ▾ Search Security ▾       

File Browser

Search for file name Actions ▾  Move to trash ▾

 Home / user / cloudera  History  Trash

| <input type="checkbox"/> | Name | Size | User | Group | Permissions | Date |
|--------------------------|---|------|----------|------------|-------------|------------------------|
| <input type="checkbox"/> |   | | hdfs | supergroup | drwxr-xr-x | June 13, 2016 08:34 PM |
| <input type="checkbox"/> |  . | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:21 PM |
| <input type="checkbox"/> |  input | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:20 PM |
| <input type="checkbox"/> |  output | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:21 PM |

Upload a local file to HDFS

The screenshot shows the Hue File Browser interface. At the top, there's a navigation bar with links for Query Editors, Data Browsers, Workflows, Search, and Security. Below the navigation bar, the main title is "File Browser". On the left, there's a search bar and some action buttons: "Actions", "Move to trash", and a plus sign. The current path is displayed as "/ user / cloudera / input". On the right, there are buttons for "Upload" (with a dropdown menu), "History", and "Trash". A red arrow points to the "Upload" button. Below the path, there's a table listing files in the directory. The columns are Size, User, Group, Permissions, and Date. There are two entries:

| Size | User | Group | Permissions | Date |
|------|----------|----------|-------------|------------------------|
| | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:21 PM |
| | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:20 PM |

Upload to /user/cloudera/input

Select files

or drag and drop them here

03_Suitability test.pdf

99% from 0.3MB x

HUE Home Query Editors Data Browsers Workflows Search Security

File Browser

Search for file name Actions Move to trash

Home / user / cloudera / input

History Trash

| | Name | Size | User | Group | Permissions | Date |
|--------------------------|-------------------------|----------|----------|----------|-------------|------------------------|
| <input type="checkbox"/> | .. | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:21 PM |
| <input type="checkbox"/> | . | | cloudera | cloudera | drwxr-xr-x | June 13, 2016 09:22 PM |
| <input type="checkbox"/> | 03_Suitability test.pdf | 336.8 KB | cloudera | cloudera | -rw-r--r-- | June 13, 2016 09:22 PM |

Hands-On: Connect to a master node via SSH

SSH Login to a master node

```
THANACHARTs-MacBook-Air:elastic-mapreduce-cli THANACHART$ ssh -i "imchadoop.pem" ub  
untu@ec2-54-201-147-59.us-west-2.compute.amazonaws.com  
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.13.0-74-generic x86_64)
```

* Documentation: <https://help.ubuntu.com/>

System information as of Sun Mar 27 09:08:01 UTC 2016

| | |
|------------------------------|-----------------------------------|
| System load: 0.0 | Processes: 135 |
| Usage of /: 27.6% of 29.40GB | Users logged in: 0 |
| Memory usage: 24% | IP address for eth0: 172.31.10.53 |
| Swap usage: 0% | |

Graph this data and manage this system at:
<https://landscape.canonical.com/>

Get cloud support with Ubuntu Advantage Cloud Guest:
<http://www.ubuntu.com/business/services/cloud>

*** System restart required ***

```
Last login: Sun Mar 27 09:08:01 2016 from node-io5.pool-125-24.dynamic.totbb.net  
ubuntu@ip-172-31-10-53:~$ █
```

Hadoop syntax for HDFS

| Command | Syntax |
|--|---|
| Listing of files in a directory | <code>hadoop fs -ls /user</code> |
| Create a new directory | <code>hadoop fs -mkdir /user/guest/newdirectory</code> |
| Copy a file from a local machine to Hadoop | <code>hadoop fs -put C:\Users\Administrator\Downloads\localfile.csv /user/rajn/newdirectory/hadoopfile.txt</code> |
| Copy a file from Hadoop to a local machine | <code>hadoop fs -get /user/rajn/newdirectory/hadoopfile.txt C:\Users\Administrator\Desktop\</code> |
| Tail last few lines of a large file in Hadoop | <code>hadoop fs -tail /user/rajn/newdirectory/hadoopfile.txt</code> |
| View the complete contents of a file in Hadoop | <code>hadoop fs -cat /user/rajn/newdirectory/hadoopfile.txt</code> |
| Remove a complete directory from Hadoop | <code>hadoop fs -rm -r /user/rajn/newdirectory</code> |
| Check the Hadoop filesystem space utilization | <code>hadoop fs -du /</code> |

Install wget

- Command: yum install wget

```
[root@quickstart /]# yum install wget
Loaded plugins: fastestmirror
Setting up Install Process
Determining fastest mirrors
epel/metalink | 13 kB     00:00
 * base: mirrors.evowise.com
 * epel: mirror.cogentco.com
 * extras: mirror.us.leaseweb.net
 * updates: mirror.cs.pitt.edu
base | 3.7 kB     00:00
base/primary_db | 4.7 MB     00:06
```

Download an example text file

Make your own directory at a master node to avoid mixing with others

```
$mkdir guest1
$cd guest1
$wget https://s3.amazonaws.com/imcbucket/input/pg2600.txt
```

```
--2016-03-27 09:58:48-- https://s3.amazonaws.com/imcbucket/input/pg2600.txt
Resolving s3.amazonaws.com (s3.amazonaws.com)... 54.231.19.187
Connecting to s3.amazonaws.com (s3.amazonaws.com)|54.231.19.187|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3291648 (3.1M) [text/plain]
Saving to: 'pg2600.txt'

100%[=====>] 3,291,648 3.14MB/s in 1.0s

2016-03-27 09:58:50 (3.14 MB/s) - 'pg2600.txt' saved [3291648/3291648]
```

Upload Data to Hadoop

Note: you login as **ubuntu**, so you need to a sudo command to Switch user to **hdfs**

```
$hadoop fs -ls /user/cloudera/input
$hadoop fs -rm /user/cloudera/input/*
$hadoop fs -put pg2600.txt /user/cloudera/input/
$hadoop fs -ls /user/cloudera/input
```

```
[root@quickstart guest1]# hadoop fs -ls /user/cloudera/input
Found 1 items
-rw-r--r-- 1 root cloudera 3291648 2016-06-14 04:29 /user/cloudera/input/pg2600.txt
[root@quickstart guest1]#
```



Lecture

Understanding Spark

Introduction

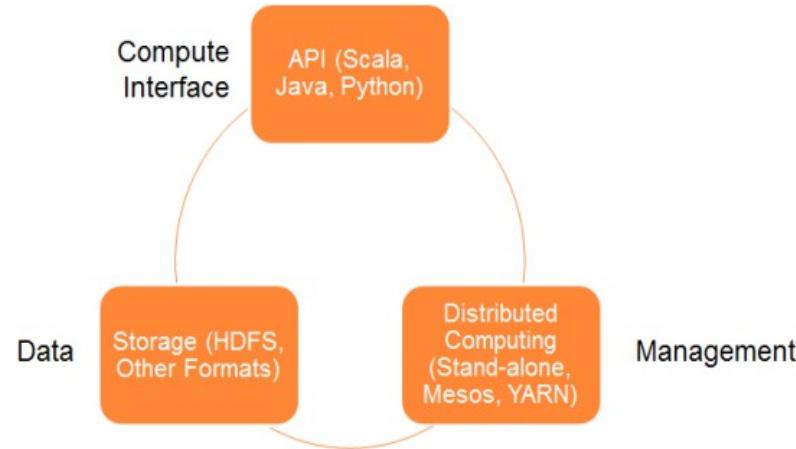
A fast and general engine for large scale data processing



An open source big data processing framework built around speed, ease of use, and sophisticated analytics. Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk.

What is Spark?

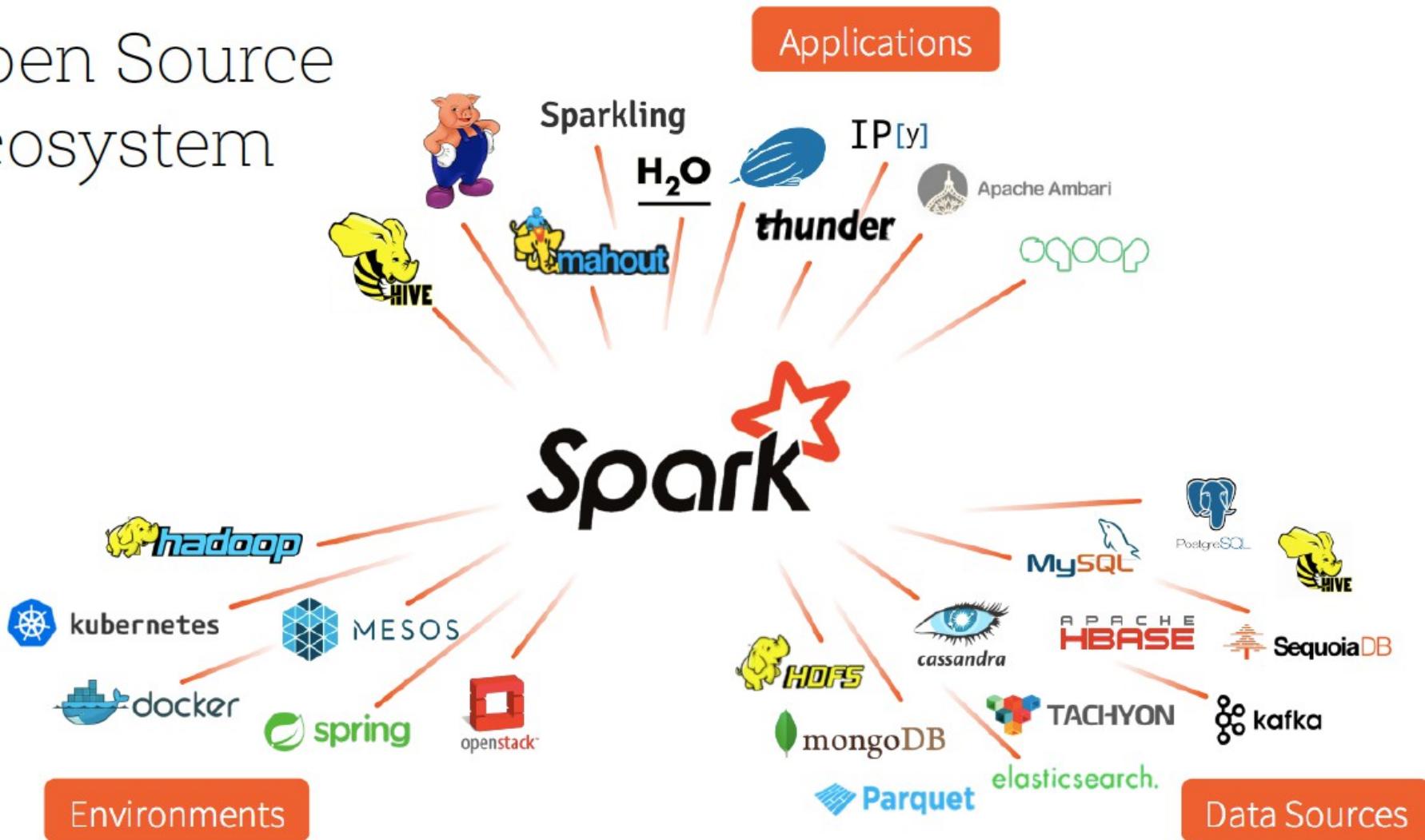
- Framework for distributed processing.
- In-memory, fault tolerant data structures
- Flexible APIs in Scala, Java, Python, SQL, R
- Open source



Why Spark?

- Handle Petabytes of data
- Significant faster than MapReduce
- Simple and intuitive APIs
- General framework
 - Runs anywhere
 - Handles (most) any I/O
 - Interoperable libraries for specific use-cases

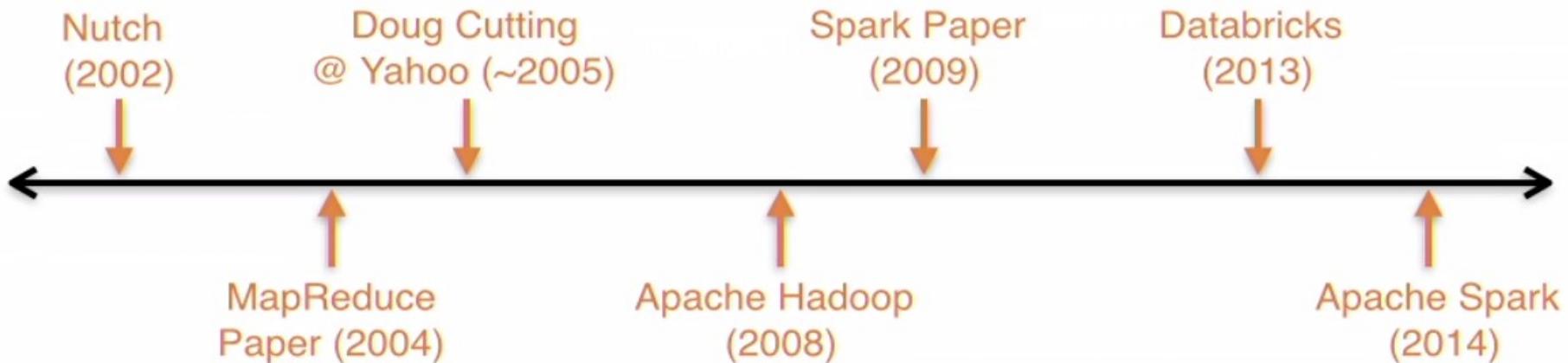
Open Source Ecosystem



Source: Jump start into Apache Spark and Databricks

Spark: History

- Founded by AMPIlab, UC Berkeley
- Created by Matei Zaharia (PhD Thesis)
- Maintained by Apache Software Foundation
- Commercial support by Databricks





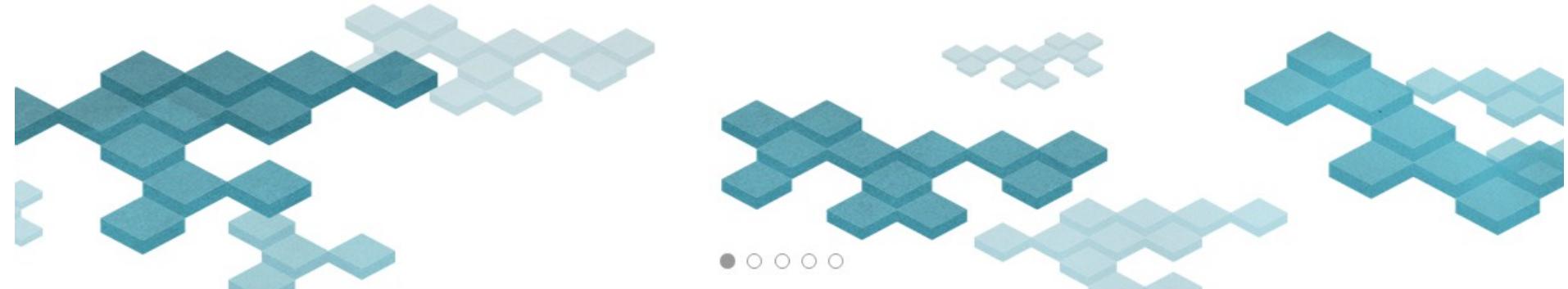
PRODUCT SPARK SOLUTIONS CUSTOMERS COMPANY BLOG RESOURCES

Partners Training [Sign Up](#) 



Data Science made easy, from ingest to production. Powered by Apache Spark™.

[SIGN UP FOR A 14-DAY FREE TRIAL](#)

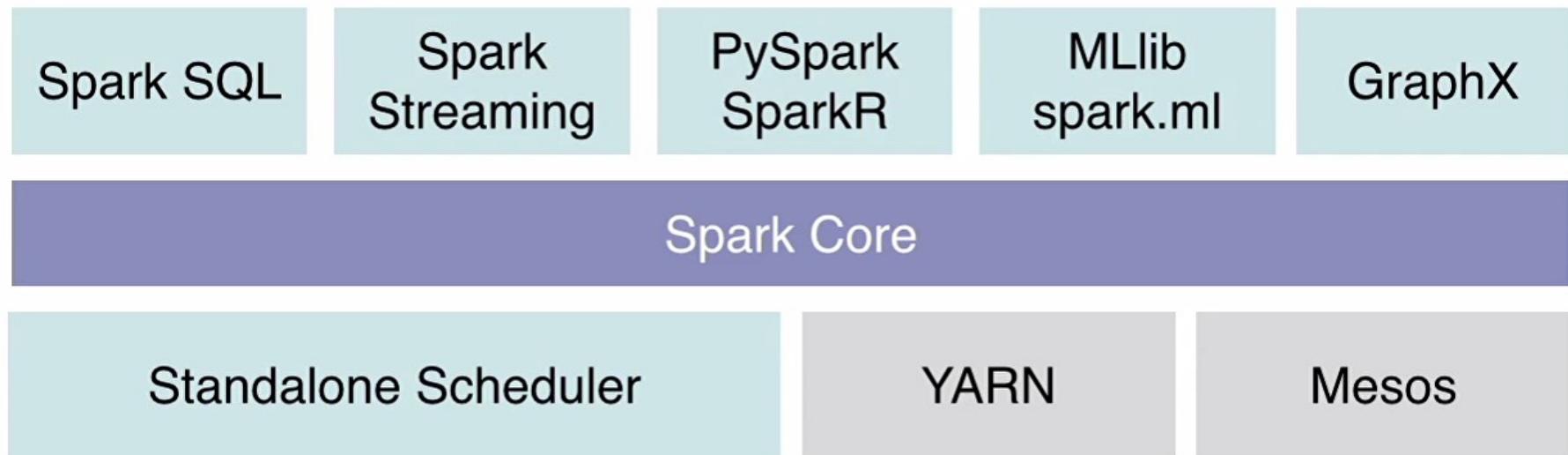


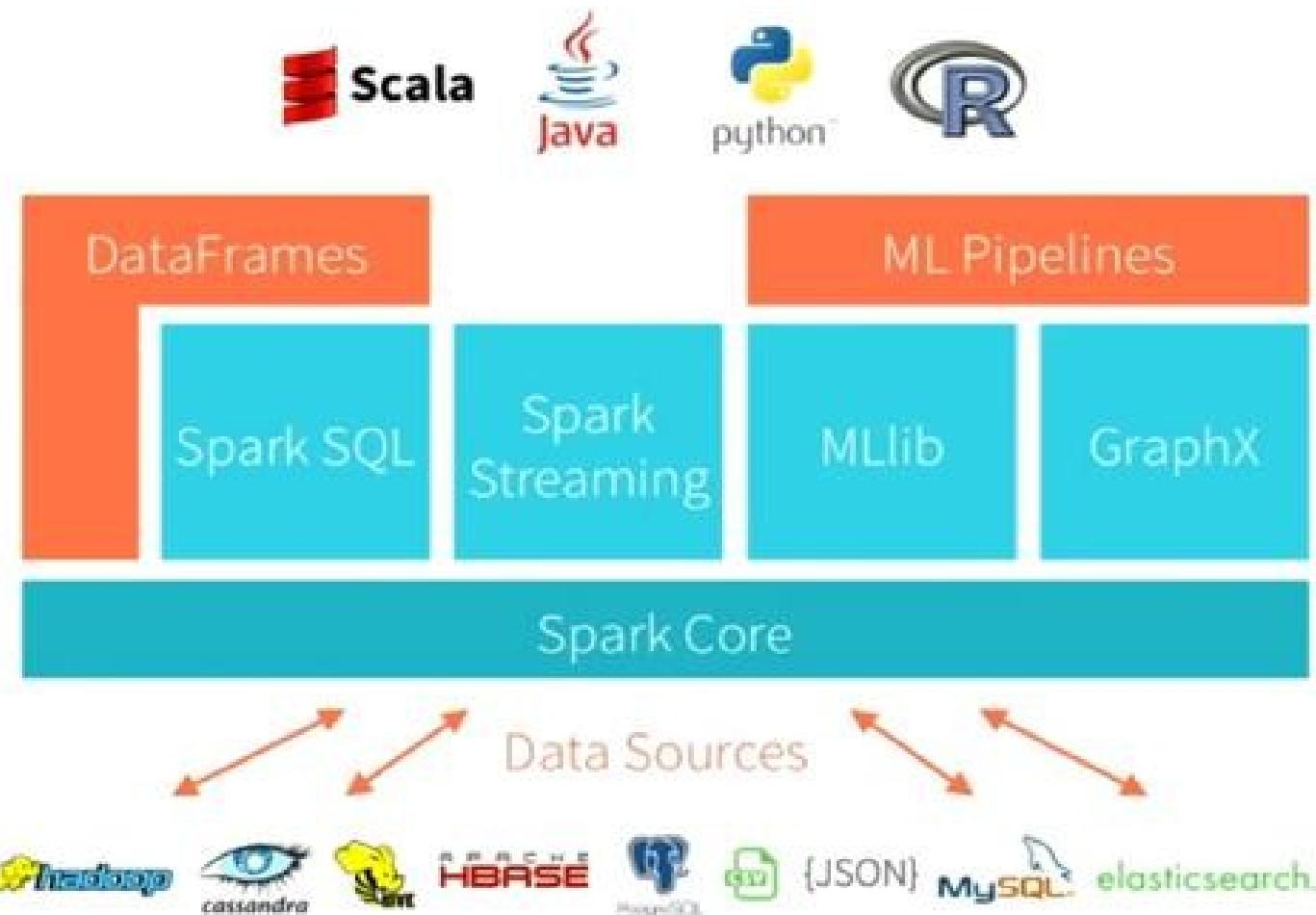
LEARN SPARK

Join the Community Edition Beta waitlist >

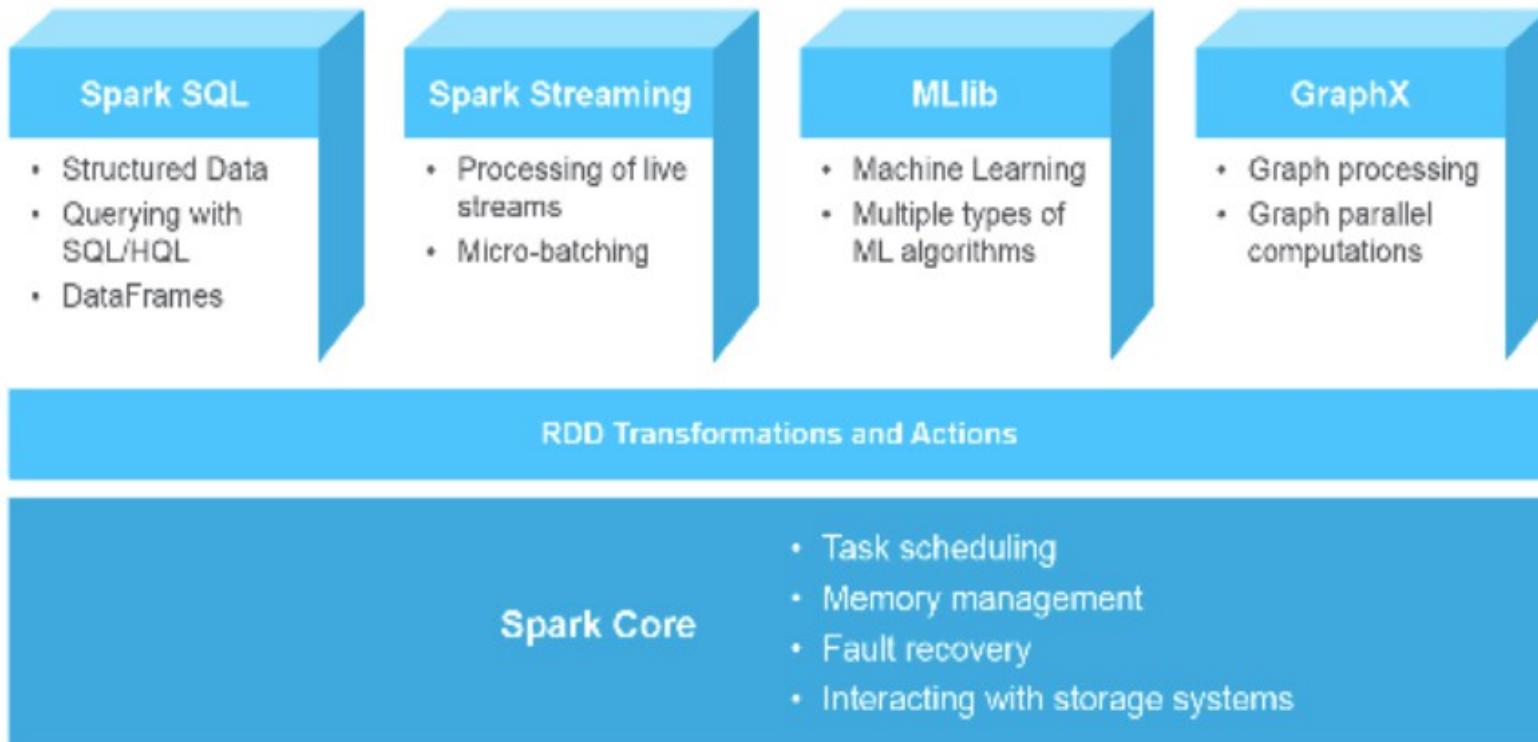
[Community Edition](#) [Spark 1.6](#) [Apache Spark 1.6](#)

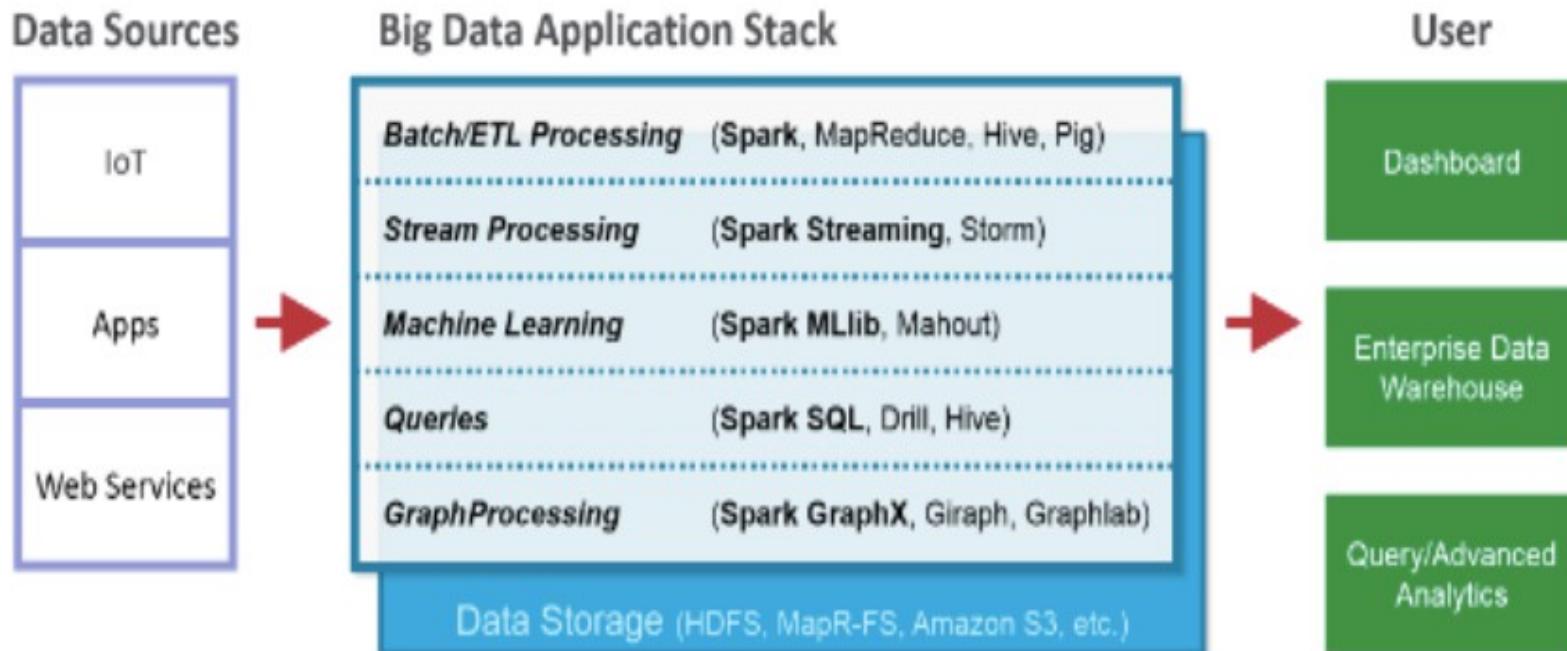
Spark Platform

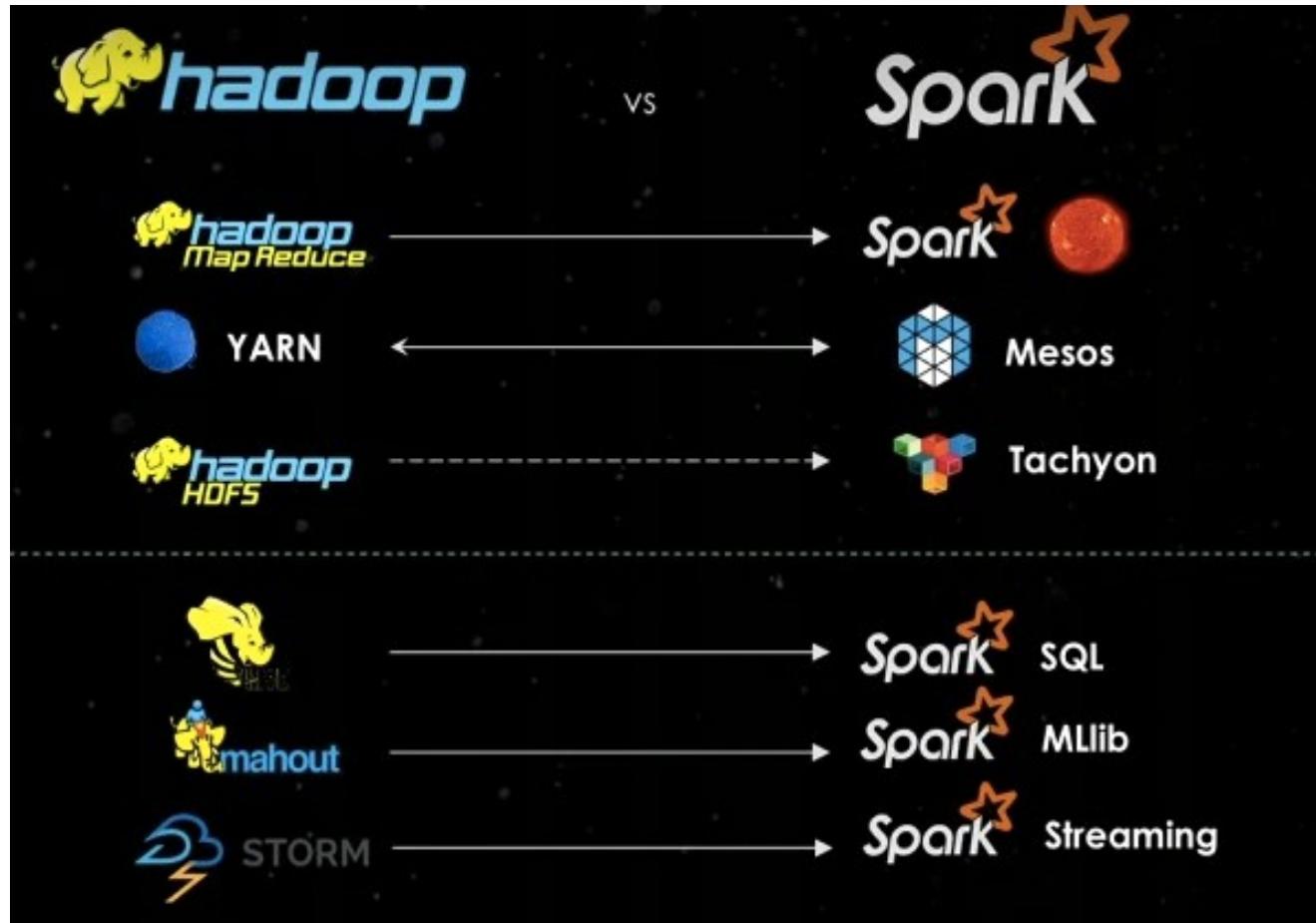




Spark Platform







Source: TRAINING Intro to Apache Spark - Brian Clapper

Large-Scale Usage



Largest cluster
8000 Nodes (Tencent)



Largest single job
1 PB (Alibaba, Databricks)



Top Streaming Intake
1 TB/hour (HHMI
Janelia Farm)



2014 On-Disk Sort Record
Fastest Open Source Engine
for sorting a PB

Notable Users

Companies That Presented at Spark Summit 2015 in San Francisco

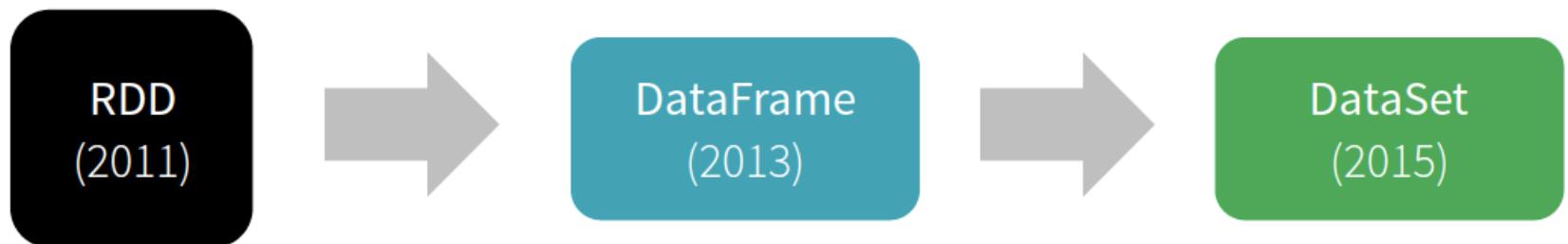


THOMSON REUTERS



Source: Jump start into Apache Spark and Databricks

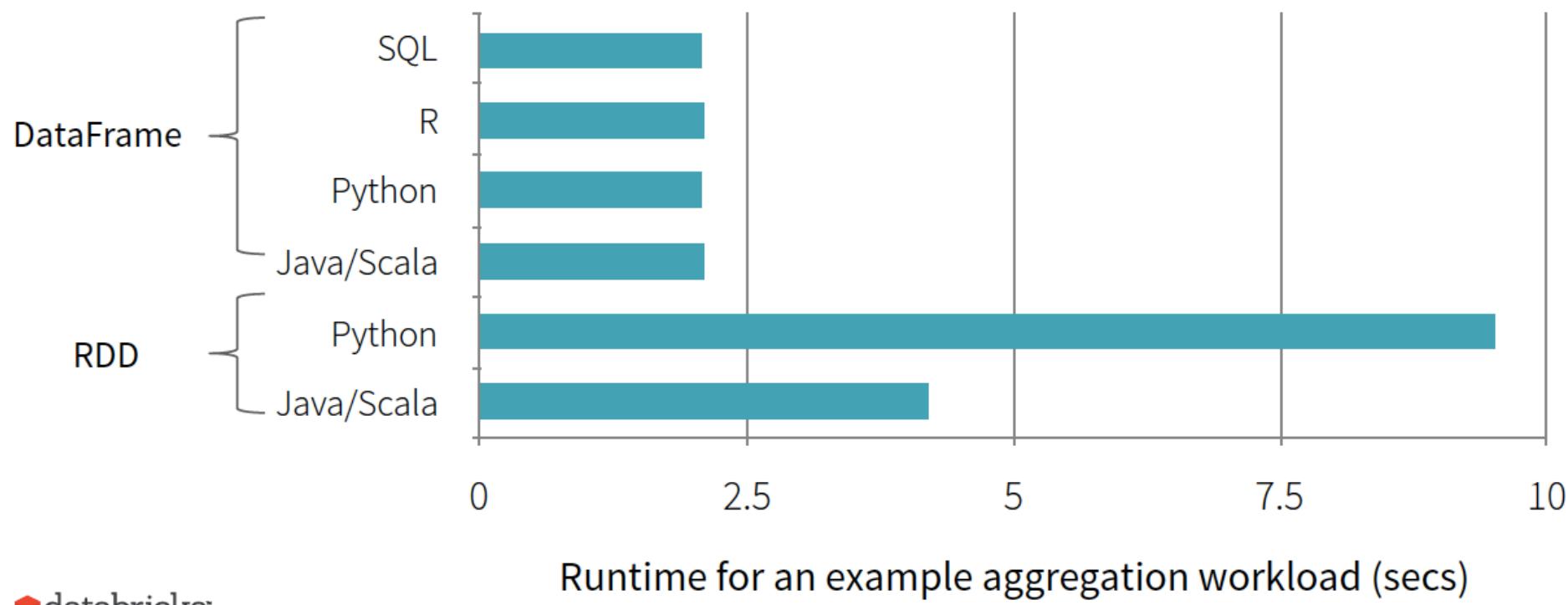
History of Spark APIs



- | | | |
|---|---|---|
| <ul style="list-style-type: none">• Distribute collection of JVM objects• Functional Operators (map, filter, etc.) | <ul style="list-style-type: none">• Distribute collection of Row objects• Expression-based operations and UDFs• Logical plans and optimizer• Fast/efficient internal representations | <ul style="list-style-type: none">• Internally rows, externally JVM objects• “Best of both worlds” type safe + fast |
|---|---|---|

Source: Jump start into Apache Spark and Databricks

Benefit of Logical Plan: Performance Parity Across Languages



Source: Jump start into Apache Spark and Databricks

What is a RDD?

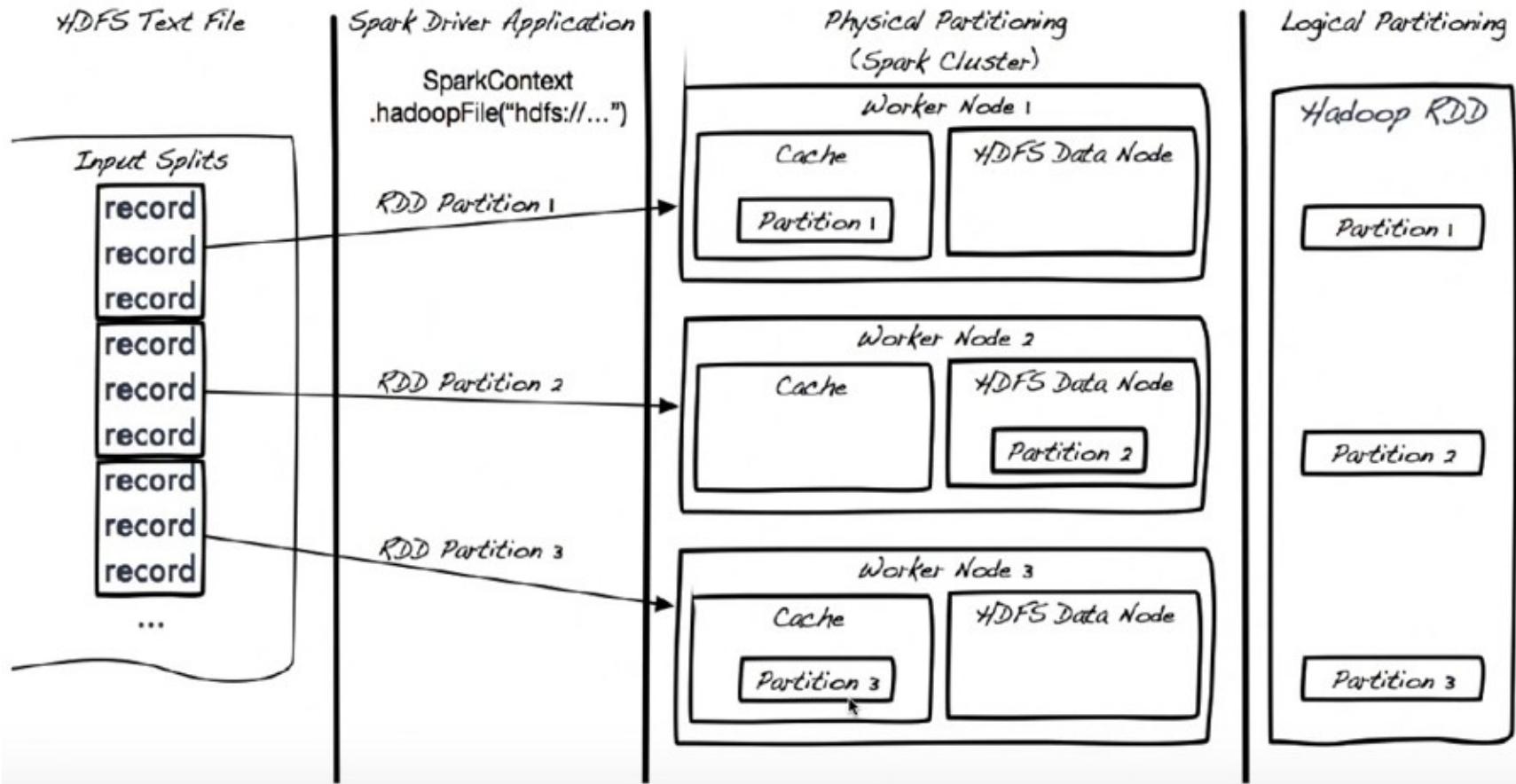
- **Resilient:** if the data in memory (or on a node) is lost, it can be recreated.
- **Distributed:** data is chunked into partitions and stored in memory across the cluster.
- **Dataset:** initial data can come from a table or be created programmatically

RDD:

- Fault tollerant
- Immutable
- Three methods for creating RDD:
 - Parallelizing an existing correction
 - Referencing a dataset
 - Transformation from an existing RDD
- Types of files supported:
 - Text files
 - SequenceFiles
 - Hadoop InputFormat

RDD Creation

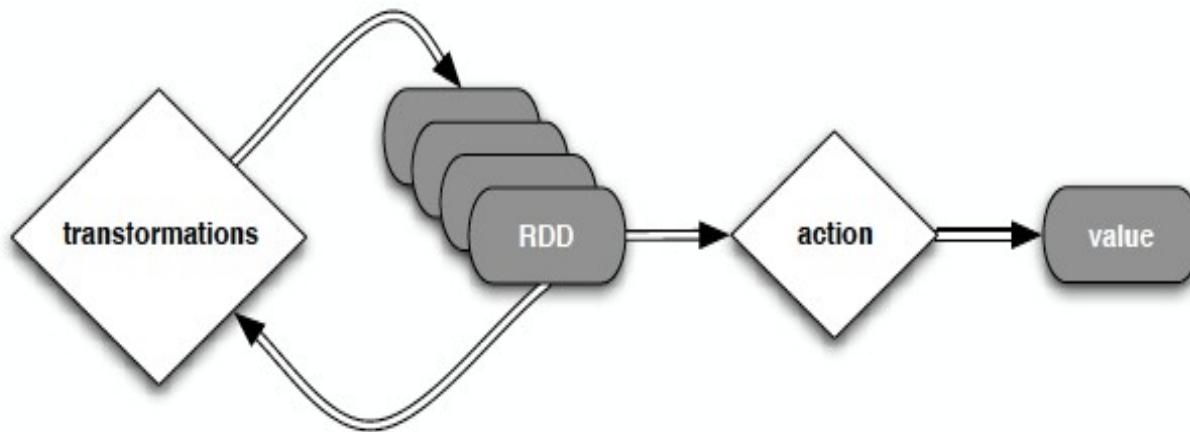
```
hdfsData = sc.textFile("hdfs://data.txt")
```



Source: Pspark: A brain-friendly introduction

RDD: Operations

- **Transformations:** transformations are lazy (not computed immediately)
- **Actions:** the transformed RDD gets recomputed when an action is run on it (default)



Direct Acyclic Graph (DAG)

- View the DAG

linesLength.toDebugString

- Sample DAG

```
res5: String =  
  MappedRDD[4] at map at <console>:16 (3 partitions)  
    MappedRDD[3] at map at <console>:16 (3 partitions)  
      FilteredRDD[2] at filter at <console>:14 (3 partitions)  
        MappedRDD[1] at textFile at <console>:12 (3 partitions)  
          HadoopRDD[0] at textFile at <console>:12 (3 partitions)|
```

Functions Deconstructed

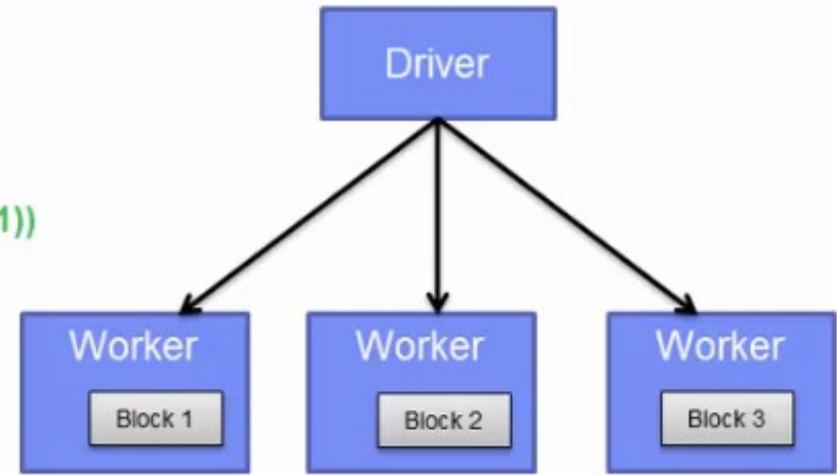
```
import random
flips = 1000000

# lazy eval
coins = xrange(flips) ← Python Generator

# lazy eval, nothing executed
heads = sc.parallelize(coins) \ ← Create RDD
Transformations → .map(lambda i: random.random()) \
    .filter(lambda r: r < 0.5) \
    .count() ← Action (materialize result)
```

What happens when an action is executed

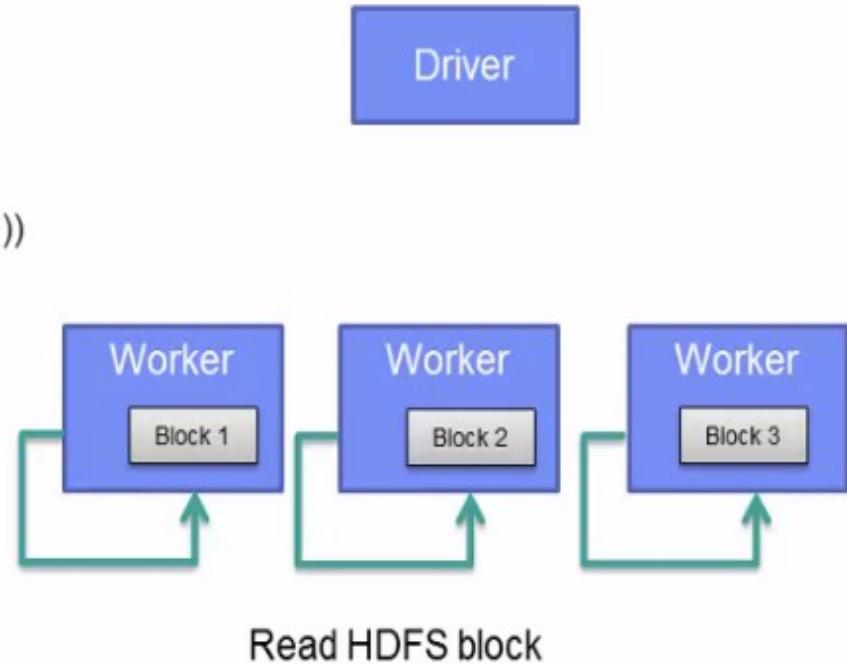
```
// Creating the RDD
val logFile = sc.textFile("hdfs://...")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
// Cache
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



Driver sends the code to be
executed on each block

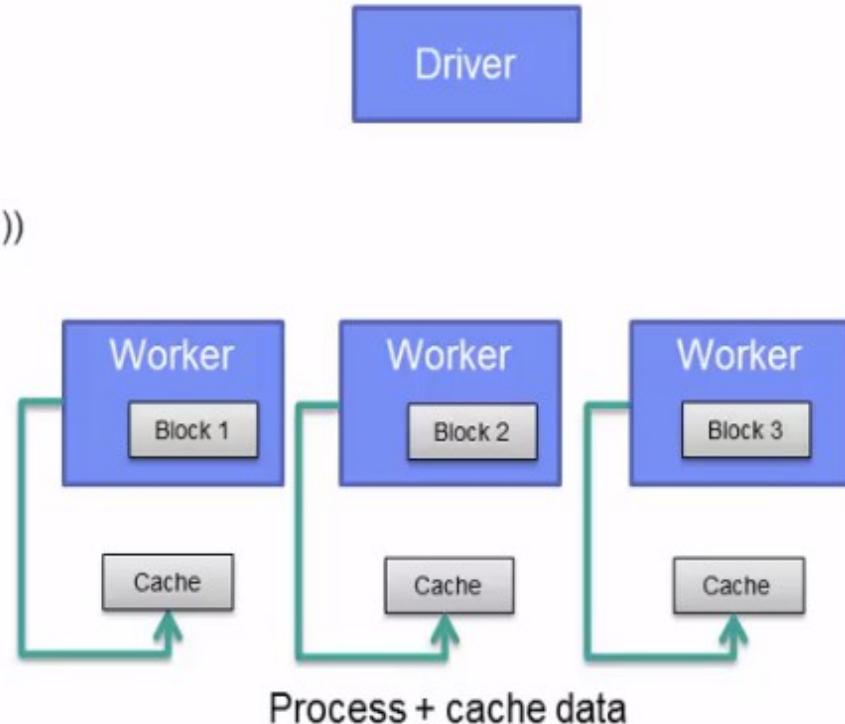
What happens when an action is executed

```
// Creating the RDD
val logFile = sc.textFile("hdfs://...")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
//Caching
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



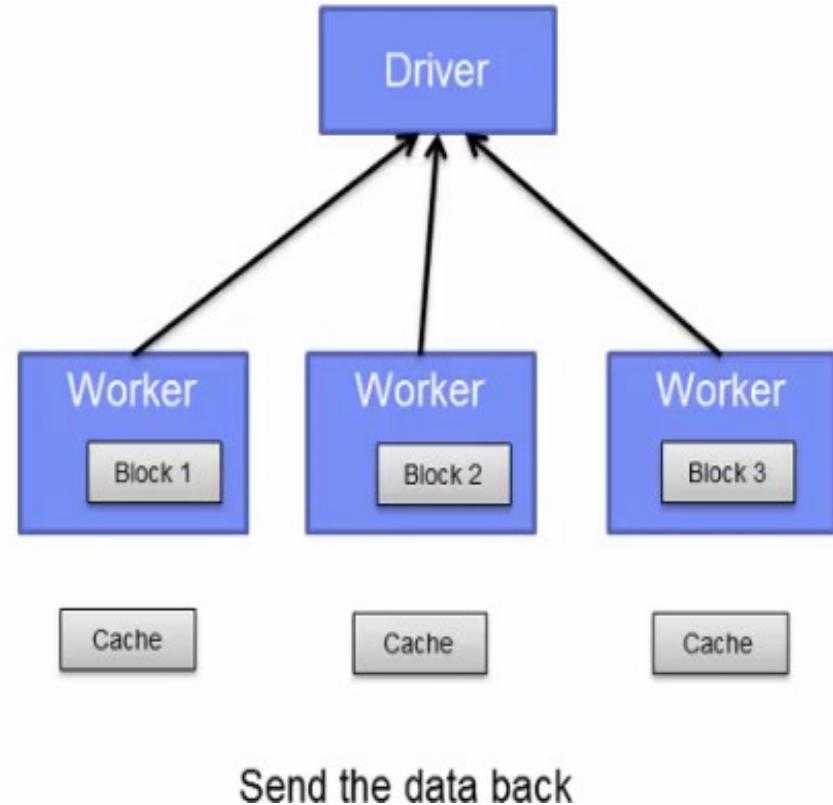
What happens when an action is executed

```
// Creating the RDD
val logFile = sc.textFile("hdfs://...")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
//Caching
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



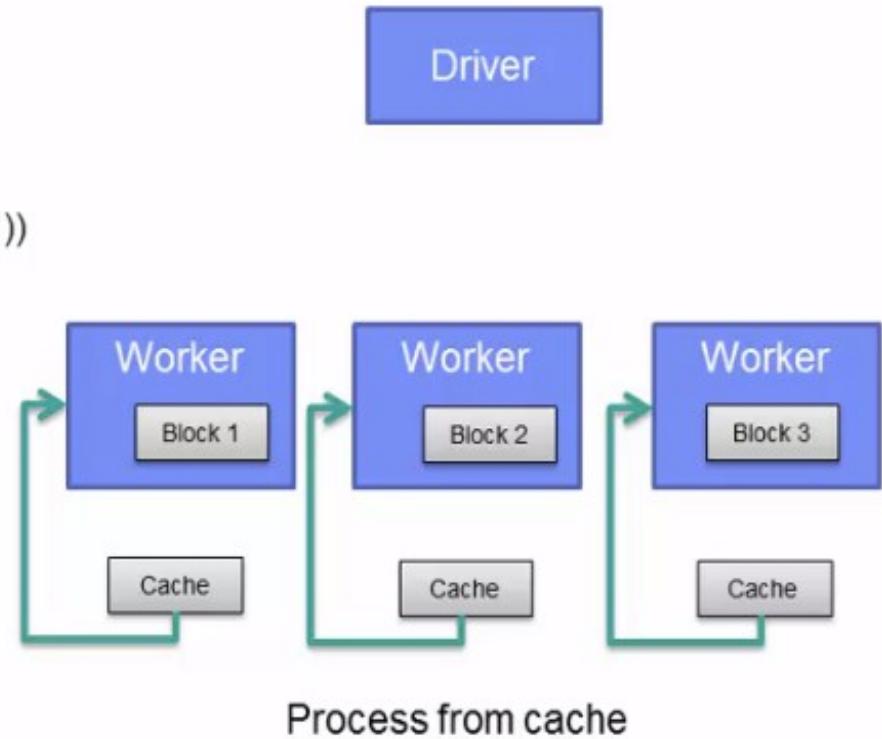
What happens when an action is executed

```
// Creating the RDD
val logFile = sc.textFile("hdfs://...")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
//Caching
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



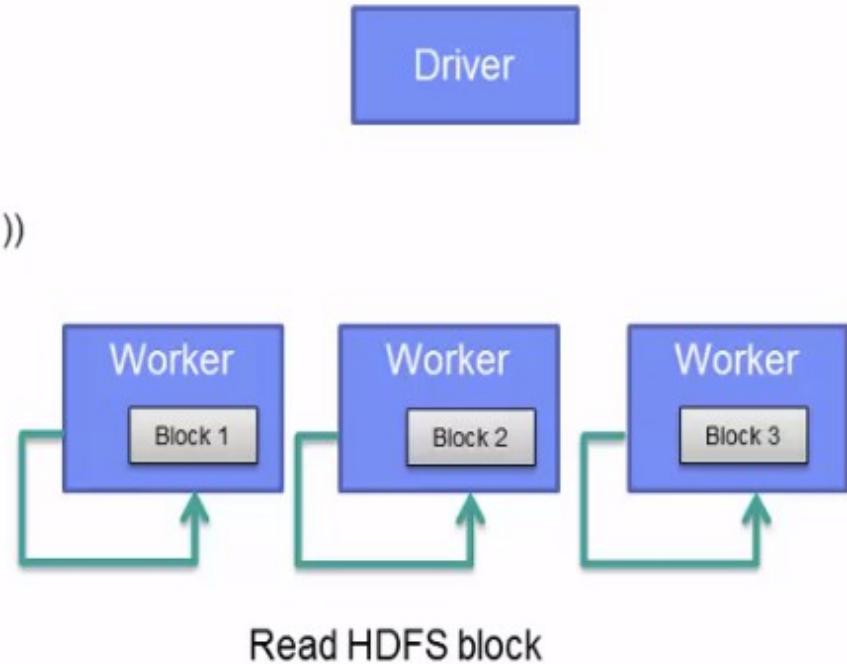
What happens when an action is executed

```
// Creating the RDD
val logFile = sc.textFile("hdfs://... ")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
//Caching
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



What happens when an action is executed

```
// Creating the RDD
val logFile = sc.textFile("hdfs://...")
// Transformations
val errors = logFile.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
//Caching
messages.cache()
// Actions
messages.filter(_.contains("mysql")).count()
messages.filter(_.contains("php")).count()
```



Spark: Transformation

| <i>transformation</i> | <i>description</i> |
|--|--|
| <code>map(func)</code> | return a new distributed dataset formed by passing each element of the source through a function <i>func</i> |
| <code>filter(func)</code> | return a new dataset formed by selecting those elements of the source on which <i>func</i> returns true |
| <code>flatMap(func)</code> | similar to map, but each input item can be mapped to 0 or more output items (so <i>func</i> should return a Seq rather than a single item) |
| <code>sample(withReplacement, fraction, seed)</code> | sample a fraction <i>fraction</i> of the data, with or without replacement, using a given random number generator <i>seed</i> |
| <code>union(otherDataset)</code> | return a new dataset that contains the union of the elements in the source dataset and the argument |
| <code>distinct([numTasks])</code> | return a new dataset that contains the distinct elements of the source dataset |

Spark: Transformation

| transformation | description |
|---|--|
| <code>groupByKey([numTasks])</code> | when called on a dataset of (K, V) pairs, returns a dataset of (K, seq[V]) pairs |
| <code>reduceByKey(func, [numTasks])</code> | when called on a dataset of (K, V) pairs, returns a dataset of (K, V) pairs where the values for each key are aggregated using the given reduce function |
| <code>sortByKey([ascending], [numTasks])</code> | when called on a dataset of (K, V) pairs where K implements Ordered, returns a dataset of (K, V) pairs sorted by keys in ascending or descending order, as specified in the boolean ascending argument |
| <code>join(otherDataset, [numTasks])</code> | when called on datasets of type (K, V) and (K, W), returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key |
| <code>cogroup(otherDataset, [numTasks])</code> | when called on datasets of type (K, V) and (K, W), returns a dataset of (K, seq[V], Seq[W]) tuples – also called groupWith |
| <code>cartesian(otherDataset)</code> | when called on datasets of types T and U, returns a dataset of (T, U) pairs (all pairs of elements) |

Single RDD Transformation

filter females to analyze female buying patterns

male1, male2, female1 -> female1

map squared values

2, 5, 6 -> 4, 25, 36

flatMap to break up a sentence into words

my name is ray -> my, name, is, ray

find the **distinct** values in a dataset

apple, apple, banana -> apple, banana

sample two values at random

apple, banana, guava -> banana, apple

Multiple RDD Transformation

union

apple, orange, banana, guava,
banana, pear

intersection

banana

subtract anything shown in Dataset B
from Dataset A

apple, orange

cartesian (every possible pair combo)

(apple, guava), (apple, banana), ...

Dataset A

apple
orange
banana

Dataset B

guava
banana
pear

Pair RDD Transformation

- reduceByKey
- groupByKey
- combineByKey
- mapValues
- flatMapValues
- keys
- values
- subtractByKey
- join
- rightOuterJoin
- leftOuterJoin
- cogroup
- sortByKey

Spark:Actions

| action | description |
|--|---|
| <code>reduce(func)</code> | aggregate the elements of the dataset using a function <code>func</code> (which takes two arguments and returns one), and should also be commutative and associative so that it can be computed correctly in parallel |
| <code>collect()</code> | return all the elements of the dataset as an array at the driver program – usually useful after a filter or other operation that returns a sufficiently small subset of the data |
| <code>count()</code> | return the number of elements in the dataset |
| <code>first()</code> | return the first element of the dataset – similar to <code>take(1)</code> |
| <code>take(n)</code> | return an array with the first <code>n</code> elements of the dataset – currently not executed in parallel, instead the driver program computes all the elements |
| <code>takeSample(withReplacement, fraction, seed)</code> | return an array with a random sample of <code>num</code> elements of the dataset, with or without replacement, using the given random number generator <code>seed</code> |

Spark:Actions

| action | description |
|---------------------------------------|---|
| <code>saveAsTextFile(path)</code> | write the elements of the dataset as a text file (or set of text files) in a given directory in the local filesystem, HDFS or any other Hadoop-supported file system. Spark will call <code>toString</code> on each element to convert it to a line of text in the file |
| <code>saveAsSequenceFile(path)</code> | write the elements of the dataset as a Hadoop SequenceFile in a given path in the local filesystem, HDFS or any other Hadoop-supported file system. Only available on RDDs of key-value pairs that either implement Hadoop's <code>Writable</code> interface or are implicitly convertible to <code>Writable</code> (Spark includes conversions for basic types like <code>Int</code> , <code>Double</code> , <code>String</code> , etc). |
| <code>countByKey()</code> | only available on RDDs of type <code>(K, V)</code> . Returns a 'Map' of <code>(K, Int)</code> pairs with the count of each key |
| <code>foreach(func)</code> | run a function <code>func</code> on each element of the dataset – usually done for side effects such as updating an accumulator variable or interacting with external storage systems |

Spark: Persistence

| <i>transformation</i> | <i>description</i> |
|--|---|
| MEMORY_ONLY | Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, some partitions will not be cached and will be recomputed on the fly each time they're needed. This is the default level. |
| MEMORY_AND_DISK | Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, store the partitions that don't fit on disk, and read them from there when they're needed. |
| MEMORY_ONLY_SER | Store RDD as serialized Java objects (one byte array per partition). This is generally more space-efficient than deserialized objects, especially when using a fast serializer, but more CPU-intensive to read. |
| MEMORY_AND_DISK_SER | Similar to MEMORY_ONLY_SER, but spill partitions that don't fit in memory to disk instead of recomputing them on the fly each time they're needed. |
| DISK_ONLY | Store the RDD partitions only on disk. |
| MEMORY_ONLY_2, MEMORY_AND_DISK_2, etc | Same as the levels above, but replicate each partition on two cluster nodes. |

Accumulators

- Similar to a MapReduce “Counter”
- A global variable to track metrics about your Spark program for debugging.
- Reasoning: Executors do not communicate with each other.
- Sent back to driver

Broadcast Variables

- Similar to a MapReduce “Distributed Cache”
- Sends read-only values to worker nodes.
- Great for lookup tables, dictionaries, etc.

Hands-On: Spark Programming

Functional tools in Python

- map
- filter
- reduce
- lambda
- IterTools
 - Chain, flatmap

Map

```
>>> a= [1,2,3]
```

```
>>> def add1(x) : return x+1
```

```
>>> map(add1, a)
```

Result: [2,3,4]

Filter

```
>>> a= [1,2,3,4]  
  
>>> def isOdd(x) : return x%2==1  
  
>>> filter(isOdd, a)  
  
Result: [1,3]
```

Reduce

```
>>> a= [1,2,3,4]  
  
>>> def add(x,y) : return x+y  
  
>>> reduce(add, a)
```

Result: 10

lambda

```
>>> (lambda x: x + 1)(3)
```

Result: 4

```
>>> map((lambda x: x + 1), [1,2,3])
```

Result: [2,3,4]

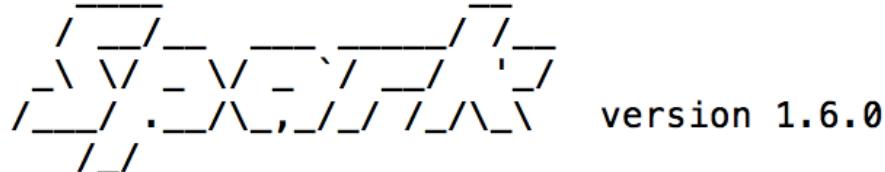
Exercises

- `(lambda x: 2*x)(3) => ?`
- `map(lambda x: 2*x, [1,2,3]) =>`
- `map(lambda t: t[0], [(1,2), (3,4), (5,6)]) =>`
- `reduce(lambda x,y: x+y, [1,2,3]) =>`
- `reduce(lambda x,y: x+y, map(lambda t: t[0], [(1,2), (3,4), (5,6)]))=>`

Start Spark-shell

```
$spark-shell
```

```
[root@quickstart 201402_babs_open_data]# spark-shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
Welcome to
```



```
Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type 'help' for more information.
```

Testing SparkContext

Spark-context

```
scala> sc
```

```
scala> sc
res0: org.apache.spark.SparkContext = org.apache.spark.SparkContext@18c07e25
```

Spark Program in Scala: WordCount

```
scala> val file =  
sc.textFile("hdfs:///user/cloudera/input/pg2600.txt")
```

```
scala> val wc = file.flatMap(l => l.split(" ")).map(word =>  
(word, 1)).reduceByKey(_ + _)
```

```
scala>  
wc.saveAsTextFile("hdfs:///user/cloudera/output/wordcountScala  
")
```

HUE [Query Editors](#) ▾ [Data Browsers](#) ▾ [Workflows](#) ▾ [Search](#) [Security](#) ▾

[File Browser](#)

Search for file name Actions ▾ Move to trash ▾

[Home](#) / user / cloudera / output / **wordcountScala** [Edit](#) History Trash

| Name | Size | User | Group | Permissions | Date |
|------------|----------|----------|----------|-------------|------------------------|
| ↑ | | cloudera | cloudera | drwxr-xr-x | June 14, 2016 02:05 AM |
| . | | root | cloudera | drwxr-xr-x | June 14, 2016 02:05 AM |
| _SUCCESS | 0 bytes | root | cloudera | -rw-r--r-- | June 14, 2016 02:05 AM |
| part-00000 | 266.3 KB | root | cloudera | -rw-r--r-- | June 14, 2016 02:05 AM |
| part-00001 | 272.7 KB | root | cloudera | -rw-r--r-- | June 14, 2016 02:05 AM |

WordCount output

HUE Home Query Editors Data Browsers Workflows Search Security

File Browser

ACTIONS

Home Page 1 of 67

View as binary

Download

View file location

Refresh

INFO

Last modified June 14, 2016 2:05 a.m.

User root

Group cloudera

Size 266.3 KB

Mode 100644

/ user / cloudera / output / wordcountScala / part-00000

```
(Ermolov.,2)
(mattered,2)
(Ah!,5)
(Koko,1)
(reunion,2)
(denied?",1)
(muslin,,1)
(intimately,3)
(blandly,5)
("Ho!,1)
(wobbers,1)
.lost...,1)
(fought?,1)
(signal.,1)
(Chem,3)
(Friend,1)
(think,",3)
(wasn't,5)
(Fve 1)
```

Spark Program in Python: WordCount

```
$ pyspark
>>> from operator import add
>>> file =
sc.textFile("hdfs:///user/cloudera/input/pg2600.txt")
>>> wc = file.flatMap(lambda x: x.split(' ')).map(lambda x:
(x, 1)).reduceByKey(add)
>>> wc.saveAsTextFile("hdfs:///user/cloudera/output/
wordcountPython")
```

File Browser

| Name | Size | User | Group | Permissions | Date |
|-----------------|------|--------|--------|-------------|---------------------------|
| wordcountPython | | guest1 | guest1 | drwxrwxrwx | January 23, 2016 11:24 PM |
| . | | ubuntu | guest1 | drwxr-xr-x | January 23, 2016 11:32 PM |
| wordcountScala | | ubuntu | guest1 | drwxr-xr-x | January 23, 2016 11:32 PM |
| | | ubuntu | guest1 | drwxr-xr-x | January 23, 2016 11:24 PM |

Project: Flight

Flight Details Data

http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

United States Department of Transportation [About DOT](#) | [Briefing Room](#) | [Our Activities](#)

OFFICE OF THE ASSISTANT SECRETARY FOR RESEARCH AND TECHNOLOGY [About OST-R](#) | [Press Room](#) | [Programs](#) | [OST-R Publications](#) | [Library](#) | [Contact Us](#)

Bureau of Transportation Statistics

[About BTS](#) | [BTS Press Room](#) | [Data and Statistics](#) | [Publications](#) | [Subject Areas](#) | [External Links](#)

[OST-R](#) > [BTS](#)

TranStats

Search this site: Advanced Search

Resources

[Database Directory](#)
[Glossary](#)
[Upcoming Releases](#)
[Data Release History](#)

Data Tools

[Analysis](#)
[Table Profile](#)
[Table Contents](#)

On-Time Performance

[Data Tables](#) [Table Contents](#)

[Download Instructions](#) [Filter Geography](#) [Filter Year](#) [Filter Period](#)

Latest Available Data: November 2015

Prezipped File % Missing Documentation Terms

| Field Name | Description | Support Table |
|--|---|----------------------------------|
| Time Period | | |
| <input type="checkbox"/> Year | Year | Get Lookup Table |
| <input type="checkbox"/> Quarter | Quarter (1-4) | Get Lookup Table |
| <input type="checkbox"/> Month | Month | Get Lookup Table |
| <input type="checkbox"/> DayofMonth | Day of Month | Get Lookup Table |
| <input type="checkbox"/> DayOfWeek | Day of Week | Get Lookup Table |
| <input type="checkbox"/> FlightDate | Flight Date (yyyymmdd) | |
| Airline | | |
| <input type="checkbox"/> UniqueCarrier | Unique Carrier Code. When the same code has been used by multiple | Get Lookup Table |

Flight Details Data

<http://stat-computing.org/dataexpo/2009/the-data.html>



ASA Sections on:

[Statistical Computing](#)
[Statistical Graphics](#)

[[Computing, Graphics](#)]

[[Awards, Data expo, Video library](#)]

[[Events, News, Newsletter](#)]

[Data expo '09](#)

Get the data

The data comes originally from [RITA](#) where it is [described in detail](#). You can download the data there, or from the bzipped csv files listed below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals.

Download individual years:

[1987](#), [1988](#), [1989](#), [1990](#), [1991](#), [1992](#), [1993](#), [1994](#), [1995](#), [1996](#), [1997](#), [1998](#), [1999](#), [2000](#), [2001](#),
[2002](#), [2003](#), [2004](#), [2005](#), [2006](#), [2007](#), [2008](#)

Data expo 09

- [Posters & results](#)
- [Competition description](#)
- [Download the data](#)
- [Supplemental data sources](#)
- [Using a database](#)
- [Intro to command line tools](#)

Data Description

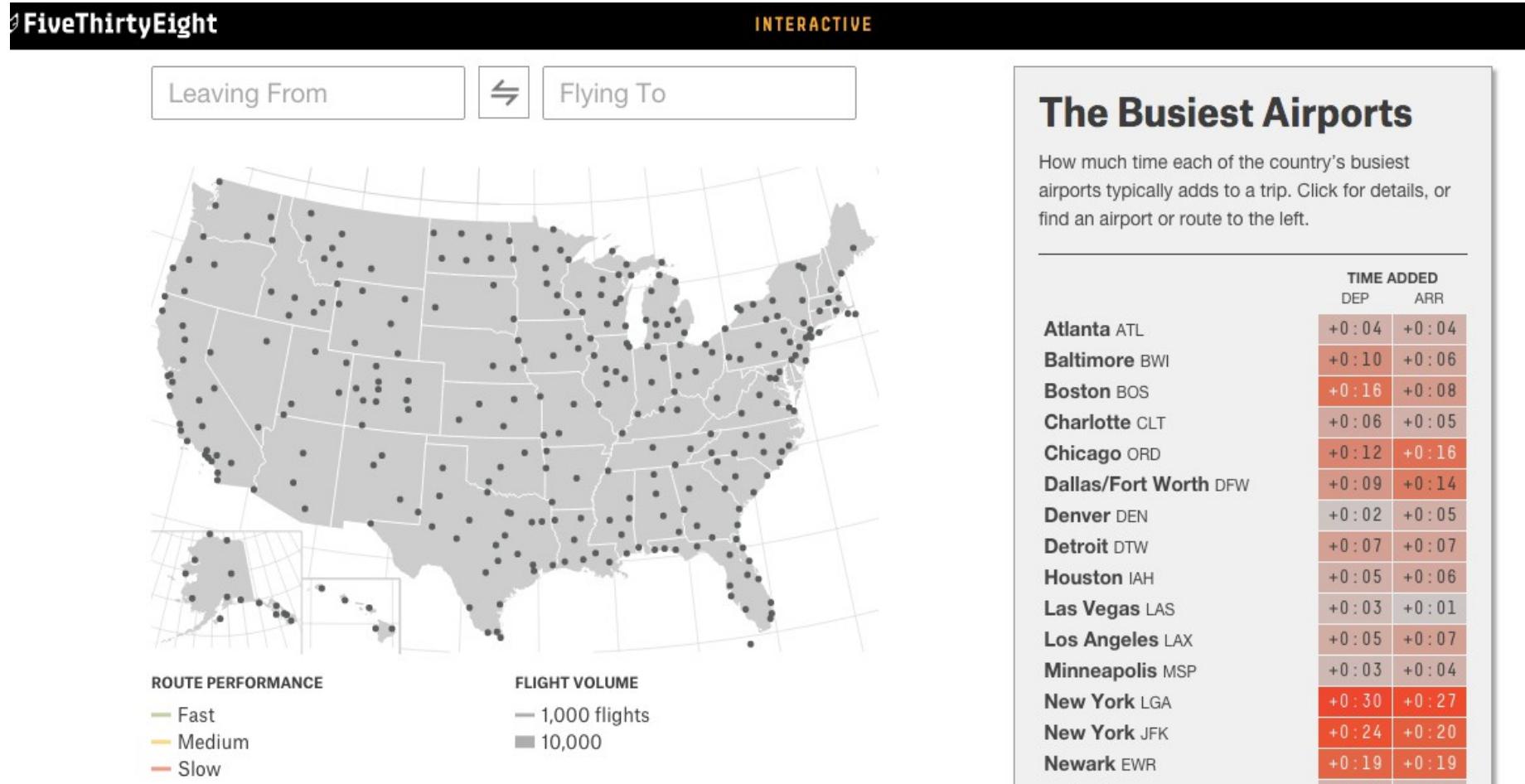
| Name | Description |
|----------------------|---|
| 1 Year | 1987-2008 |
| 2 Month | 1-12 |
| 3 DayofMonth | 1-31 |
| 4 DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 DepTime | actual departure time (local, hhmm) |
| 6 CRSDepTime | scheduled departure time (local, hhmm) |
| 7 ArrTime | actual arrival time (local, hhmm) |
| 8 CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 UniqueCarrier | <u>unique carrier code</u> |
| 10 FlightNum | flight number |
| 11 TailNum | plane tail number |
| 12 ActualElapsedTime | in minutes |
| 13 CRSElapsedTime | in minutes |
| 14 AirTime | in minutes |
| 15 ArrDelay | arrival delay, in minutes |
| 16 DepDelay | departure delay, in minutes |
| 17 Origin | origin <u>IATA airport code</u> |
| 18 Dest | destination <u>IATA airport code</u> |
| 19 Distance | in miles |
| 20 TaxiIn | taxi in time, in minutes |
| 21 TaxiOut | taxi out time in minutes |
| 22 Cancelled | was the flight cancelled? |
| 23 CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 Diverted | 1 = yes, 0 = no |
| 25 CarrierDelay | in minutes |
| 26 WeatherDelay | in minutes |
| 27 NASDelay | in minutes |
| 28 SecurityDelay | in minutes |
| 29 LateAircraftDelay | in minutes |

Snapshot of Dataset

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | |
|----|------|-------|---------|---------|---------|----------|---------|----------|----------|-----------|---------|------------|-----------|---------|----------|----------|--------|------|----------|--------|---------|
| 1 | Year | Month | DayofMo | DayOfWe | DepTime | CRSDepTl | ArrTime | CRSArrTl | UniqueCa | FlightNum | TailNum | ActualElap | CRSElapse | AirTime | ArrDelay | DepDelay | Origin | Dest | Distance | TaxiIn | TaxiOut |
| 2 | 2008 | 1 | 5 | 6 | 2243 | 1415 | 45 | 1625 | WN | 1684 | N347SW | 62 | 70 | 41 | 500 | 508 | SAN | PHX | 304 | 2 | |
| 3 | 2008 | 1 | 5 | 6 | 1940 | 1220 | 2111 | 1350 | WN | 1684 | N347SW | 91 | 90 | 64 | 441 | 440 | SFO | SAN | 447 | 5 | |
| 4 | 2008 | 1 | 7 | 1 | 111 | 1845 | 308 | 2045 | WN | 405 | N644SW | 117 | 120 | 103 | 383 | 386 | MDW | JAN | 666 | 4 | |
| 5 | 2008 | 1 | 7 | 1 | 2213 | 1700 | 2317 | 1655 | WN | 1827 | N759GS | 124 | 55 | 75 | 382 | 313 | IND | MDW | 162 | 10 | |
| 6 | 2008 | 1 | 7 | 1 | 2143 | 1720 | 26 | 1820 | WN | 1430 | N644SW | 163 | 60 | 83 | 366 | 263 | STL | MDW | 251 | 24 | |
| 7 | 2008 | 1 | 7 | 1 | 117 | 2020 | 302 | 2135 | WN | 490 | N651SW | 105 | 75 | 87 | 327 | 297 | STL | TUL | 351 | 5 | |
| 8 | 2008 | 1 | 7 | 1 | 2358 | 1855 | 105 | 2000 | WN | 490 | N651SW | 67 | 65 | 50 | 305 | 303 | MDW | STL | 251 | 4 | |
| 9 | 2008 | 1 | 3 | 4 | 2245 | 1730 | 2354 | 1850 | WN | 186 | N792SW | 69 | 80 | 59 | 304 | 315 | JAN | HOU | 359 | 3 | |
| 10 | 2008 | 1 | 7 | 1 | 2219 | 1730 | 35 | 1935 | WN | 2474 | N710SW | 76 | 65 | 67 | 300 | 289 | MDW | CMH | 284 | 2 | |
| 11 | 2008 | 1 | 5 | 6 | 2129 | 1620 | 2246 | 1750 | WN | 1924 | N408WN | 77 | 90 | 56 | 296 | 309 | SFO | LAS | 414 | 4 | |
| 12 | 2008 | 1 | 3 | 4 | 1615 | 1130 | 1623 | 1135 | WN | 10 | N617SW | 68 | 65 | 56 | 288 | 285 | MAF | ABQ | 332 | 4 | |
| 13 | 2008 | 1 | 3 | 4 | 1736 | 1305 | 2031 | 1555 | WN | 1837 | N761RR | 255 | 290 | 268 | 276 | 271 | MDW | SFO | 1855 | 4 | |
| 14 | 2008 | 1 | 5 | 6 | 2236 | 1805 | 2400 | 1930 | WN | 646 | N283WN | 84 | 85 | 71 | 270 | 271 | LAX | SFO | 337 | 6 | |
| 15 | 2008 | 1 | 3 | 4 | 2021 | 1700 | 2303 | 1835 | WN | 2005 | N302SW | 162 | 95 | 73 | 268 | 201 | LAS | SFO | 414 | 4 | |
| 16 | 2008 | 1 | 3 | 4 | 2059 | 1620 | 2216 | 1750 | WN | 1924 | N761RR | 77 | 90 | 60 | 266 | 279 | SFO | LAS | 414 | 6 | |
| 17 | 2008 | 1 | 7 | 1 | 2348 | 2105 | 307 | 2250 | WN | 3137 | N358SW | 259 | 165 | 244 | 257 | 163 | MCO | MDW | 989 | 1 | |
| 18 | 2008 | 1 | 3 | 4 | 2255 | 1820 | 509 | 55 | WN | 1924 | N761RR | 194 | 215 | 176 | 254 | 275 | LAS | IND | 1591 | 9 | |
| 19 | 2008 | 1 | 9 | 3 | 1458 | 1040 | 1725 | 1315 | WN | 2556 | N501SW | 87 | 95 | 76 | 250 | 258 | BNA | BWI | 588 | 4 | |
| 20 | 2008 | 1 | 7 | 1 | 2300 | 1835 | 113 | 2105 | WN | 2804 | N420WN | 253 | 270 | 240 | 248 | 265 | MDW | PDX | 1751 | 5 | |
| 21 | 2008 | 1 | 5 | 6 | 47 | 2040 | 151 | 2145 | WN | 505 | N435WN | 64 | 65 | 51 | 246 | 247 | BWI | PVD | 328 | 5 | |
| 22 | 2008 | 1 | 5 | 6 | 1558 | 1225 | 14 | 2010 | WN | 505 | N442WN | 316 | 285 | 250 | 244 | 213 | SAN | BWI | 2295 | 5 | |
| 23 | 2008 | 1 | 5 | 6 | 1931 | 1540 | 2104 | 1705 | WN | 1179 | N718SW | 93 | 85 | 77 | 239 | 231 | SAN | OAK | 446 | 7 | |
| 24 | 2008 | 1 | 4 | 5 | 1822 | 1425 | 2003 | 1605 | WN | 753 | N726SW | 101 | 100 | 88 | 238 | 237 | PDX | OAK | 543 | 6 | |

FiveThirtyEight

<http://projects.fivethirtyeight.com/flights/>



Spark Program : Upload Flight Delay Data

Upload a data to HDFS

```
$ wget  
https://s3.amazonaws.com/imcbucket/data/flights/2008.csv  
$ hadoop fs -put 2008.csv /user/cloudera/input
```

The screenshot shows the Hue File Browser interface. The top navigation bar includes links for Query Editors, Data Browsers, Workflows, Search, and Security. Below the navigation is a toolbar with icons for file operations like upload, download, and trash. A search bar and action buttons are also present. The main area displays a file tree under the path /user/cloudera/input. The tree shows three entries: a folder named '2008', a file named '2008.csv' (size 657.5 MB), and a file named '.' (size 0). The 'Actions' dropdown menu is open, showing options like 'Move to trash'. The bottom of the screen shows a navigation bar with Home, History, and Trash buttons.

| Name | Size | User | Group | Permissions | Date |
|----------|----------|----------|----------|-------------|------------------------|
| 2008 | | cloudera | cloudera | drwxr-xr-x | June 14, 2016 08:54 AM |
| . | | root | cloudera | drwxr-xr-x | June 14, 2016 08:56 AM |
| 2008.csv | 657.5 MB | root | cloudera | -rw-r--r-- | June 14, 2016 08:56 AM |

Spark Program : Navigating Flight Delay Data

```
>>> airline =  
sc.textFile("hdfs://user/cloudera/input/2008.csv")  
>>> airline.take(2)
```

```
[u'Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCARRIER,  
FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,  
Origin,Dest,Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverted,CARRIERDelay,  
WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay', u'2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,150,116,-14,8,IAD,TPA,810,4,8,0,,0,NA,NA,NA,  
NA,NA']
```

Spark Program : Preparing Data

```
>>> header_line = airline.first()
>>> header_list = header_line.split(',')
>>> airline_no_header = airline.filter(lambda row: row != header_line)
>>> airline_no_header.first()
>>> def make_row(row):
...     row_list = row.split(',')
...     d = dict(zip(header_list, row_list))
...     return d
...
>>> airline_rows = airline_no_header.map(make_row)
>>> airline_rows.take(5)
```

Spark Program : Define convert_float function

```
>>> def convert_float(value):  
...     try:  
...         x = float(value)  
...         return x  
...     except ValueError:  
...         return 0  
...  
>>>
```

Spark Program : Finding best/worst airline

```
>>> carrier_rdd = airline_rows.map(lambda row:  
(row['UniqueCarrier'],convert_float(row['ArrDelay'])))  
>>> carrier_rdd.take(2)
```

```
16/06/17 17:48:06 INFO scheduler.DAGScheduler: Job 5 finished: runJob at PythonRDD.s  
cala:393, took 0.062345 s  
[(u'WN', -14.0), (u'WN', 2.0)]
```

Spark Program : Finding best/worst airlines

```
>>> mean_delays_dest =  
carrier_rdd.groupByKey().mapValues(lambda delays:  
sum(delays.data)/len(delays.data))  
  
>>> mean_delays_dest.sortBy(lambda t:t[1],  
ascending=True).take(10)  
  
>>> mean_delays_dest.sortBy(lambda t:t[1],  
ascending=False).take(10)
```

```
[..., ...]  
[(u'AQ', -2.8708974358974357), (u'HA', 1.2518519716624075), (u'US', 2.80099826053982  
78), (u'9E', 3.9874908469611912), (u'AS', 4.721360405553864), (u'WN', 5.115703380225  
9031), (u'F9', 6.0841356696810847), (u'OO', 6.4389386397817896), (u'NW', 7.293465879  
6727758), (u'DL', 7.7161646357519178)]
```

```
[(u'AA', 12.202853434950445), (u'OH', 11.404110178283158), (u'YV', 11.32256697917075  
3), (u'UA', 11.001550560048052), (u'B6', 10.859381613638567), (u'CO', 10.80982057596  
6226), (u'XE', 10.320298523403915), (u'EV', 10.00033146217589), (u'MQ', 9.4969706109  
522658), (u'FL', 8.9881574723712561)]
```

Spark SQL

DataFrame

- A distributed collection of rows organized into named columns.
- An abstraction for selecting, filtering, aggregating, and plotting structured data.
- Previously => SchemaRDD

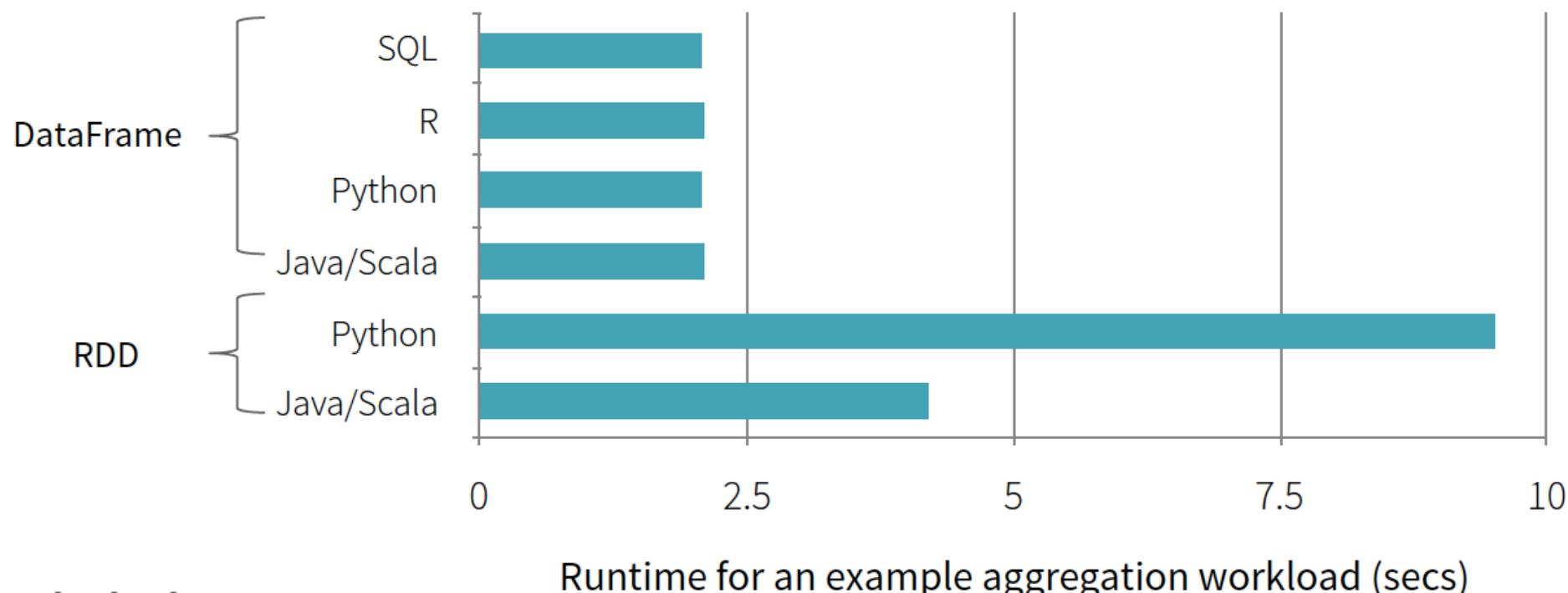
SparkSQL

- Creating and running Spark program faster
 - Write less code
 - Read less data
 - Let the optimizer do the hard work



SparkSQL
is about more than SQL.

Benefit of Logical Plan: Performance Parity Across Languages



Source: Jump start into Apache Spark and Databricks

SparkSQL

```
context = ps.HiveContext(sc)

# query with SQL
results = context.sql(
    "SELECT * FROM people")

# apply Python transformation
names = results.map(lambda p: p.name)
```

Spark SQL

Spark Core

Preparing Large Dataset

<http://grouplens.org/datasets/movielens/>



[about](#) [datasets](#) [publications](#) [blog](#)

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (<http://movielens.org>). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please [take a short survey](#) about the MovieLens datasets

MovieLens 100k

100,000 ratings from 1000 users on 1700 movies.

- [README.txt](#)
- [ml-100k.zip](#)
- [Index of unzipped files](#)

MovieLens 1M

1 million ratings from 6000 users on 4000 movies.

- [README.txt](#)

Datasets

[MovieLens](#)

[HetRec 2011](#)

[WikiLens](#)

[Book-Crossing](#)

[Jester](#)

[EachMovie](#)

MovieLen Dataset

1) Type command > `wget`

`http://files.grouplens.org/datasets/movielens/ml-100k.zip`

2) Type command > `yum install unzip`

3) Type command > `unzip ml-100k.zip`

4) Type command > `more ml-100k/u.user`

```
[root@quickstart guest1]# more ml-100k/u.user
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
11|39|F|other|30329
```

Moving dataset to HDFS

- 1) Type command > `cd ml-100k`
- 2) Type command > `hadoop fs -mkdir /user/cloudera/movielens`
- 3) Type command > `hadoop fs -put u.user /user/cloudera/movielens`
- 4) Type command > `hadoop fs -ls /user/cloudera/movielens`

```
[root@quickstart ml-100k]# hadoop fs -ls /user/cloudera/movielens
Found 1 items
-rw-r--r--  1 root cloudera      22628 2016-06-14 08:04 /user/cloudera/
movielens/u.user
[root@quickstart ml-100k]#
```

SQL Spark MovieLens

Upload a data to HDFS then

```
$ pyspark --packages com.databricks:spark-csv_2.10:1.2.0

>>> df =
sqlContext.read.format('com.databricks.spark.csv') .options(header='true') .load('hdfs://user/cloudera/u.user')

>>> df.registerTempTable('user')

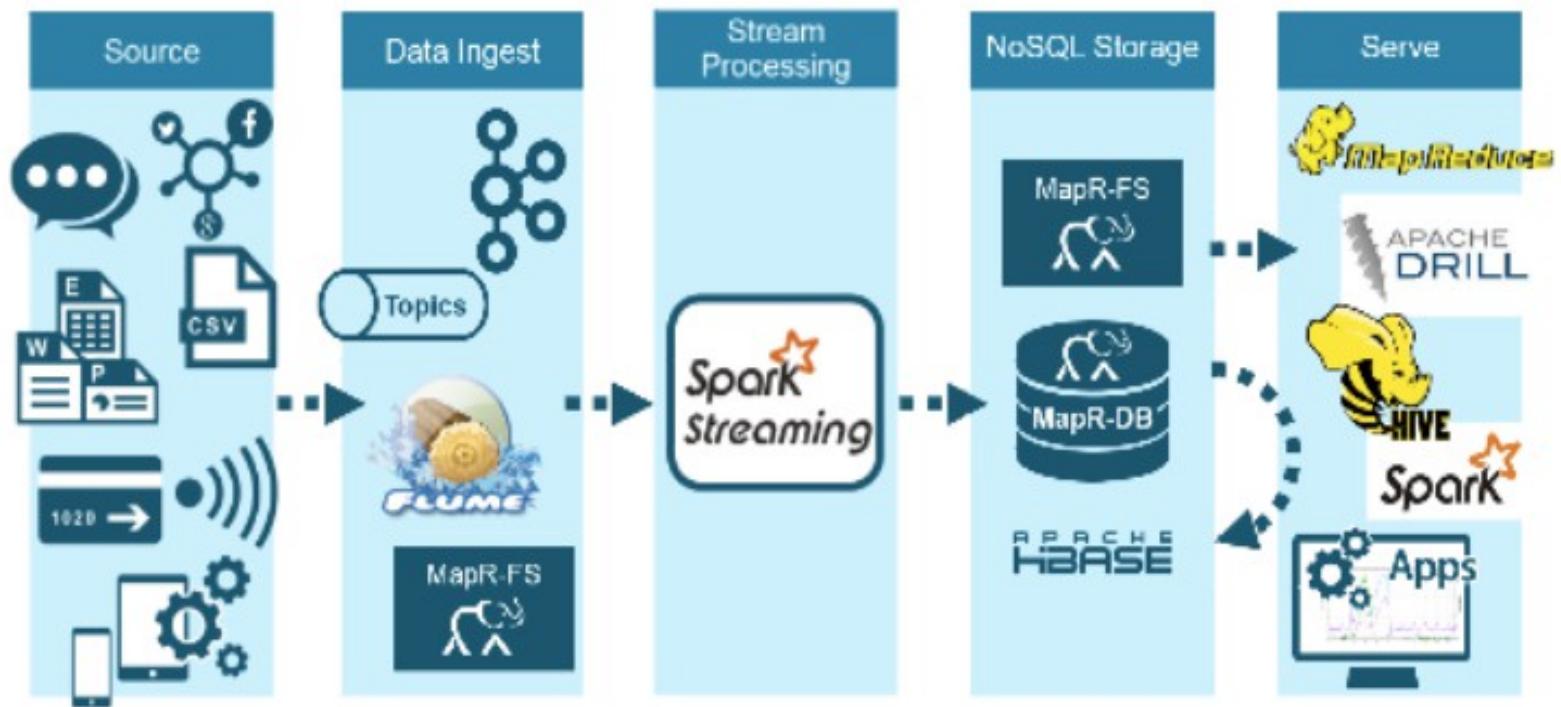
>>> sqlContext.sql("SELECT * FROM user") .collect()
```

```
16/06/17 09:47:14 INFO scheduler.DAGScheduler: Job 1 finished: collect at <stdin>:1, took 0.295085 s
```

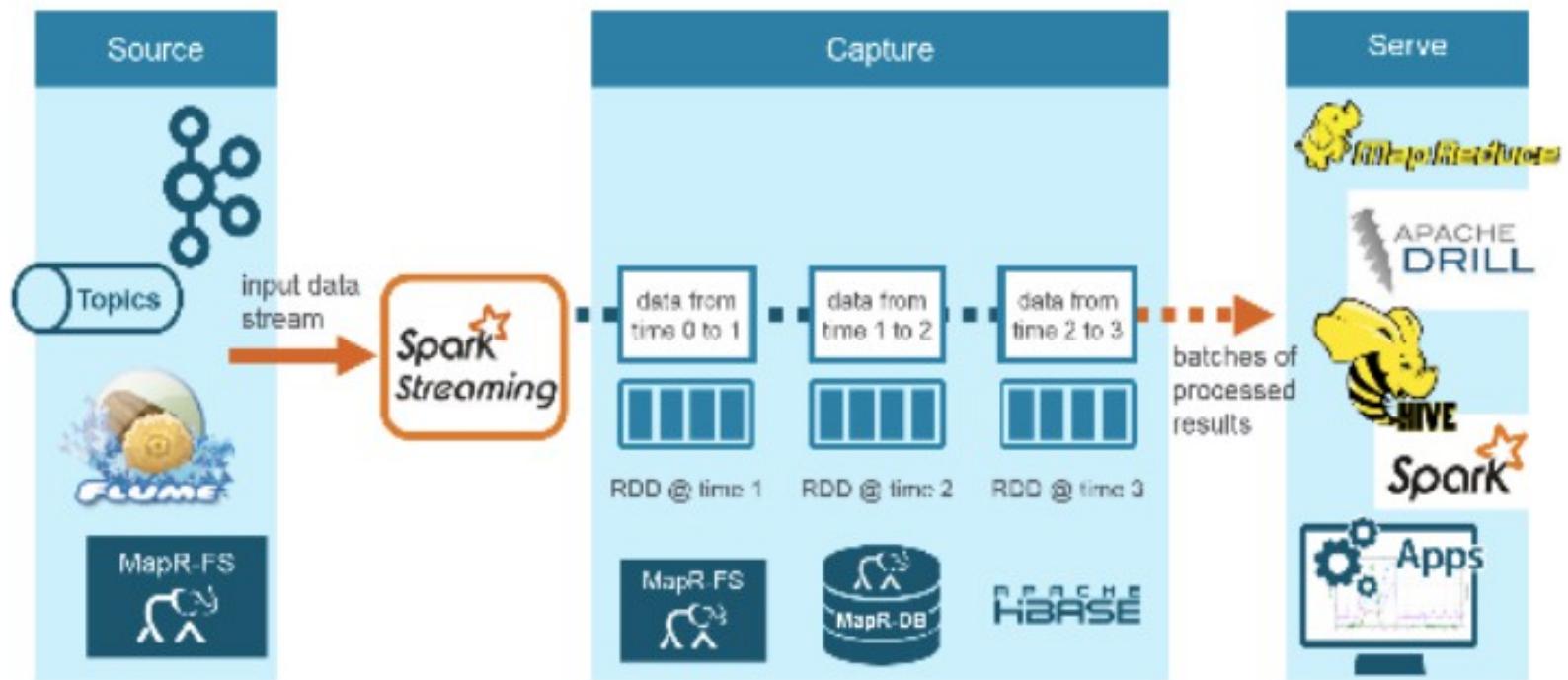
```
[Row(1|24|M|technician|85711=u'2|53|F|other|94043'), Row(1|24|M|technician|85711=u'3|23|M|writer|32067'), Row(1|24|M|technician|85711=u'4|24|M|technician|43537'), Row(1|24|M|technician|85711=u'5|33|F|other|15213'), Row(1|24|M|technician|85711=u'6|42|M|executive|98101'), Row(1|24|M|technician|85711=u'7|57|M|administrator|91344'), Row(1|24|M|technician|85711=u'8|36|M|administrator|05201'), Row(1|24|M|technician|85711=u'9|29|M|student|01002'), Row(1|24|M|technician|85711=u'10|53|M|lawyer|90703'), Row(1|24|M|technician|85711=u'11|39|F|other|30329'), Row(1|24|M|technician|85711=u'12|28|F|other|06405'), Row(1|24|M|technician|85711=u'13|47|M|educator|29206'), Row(1|24|M|technician|85711=u'14|45|M|scientist|55106'), Row(1|24|M|technician|85711=u'15|49|F|educator|97301'), Row(1|24|M|technician|85711=u'16|21|M|entertainment|10309'), Row(1|24|M|technician|85711=u'17|30|
```

Spark Streaming

Stream Process Architecture

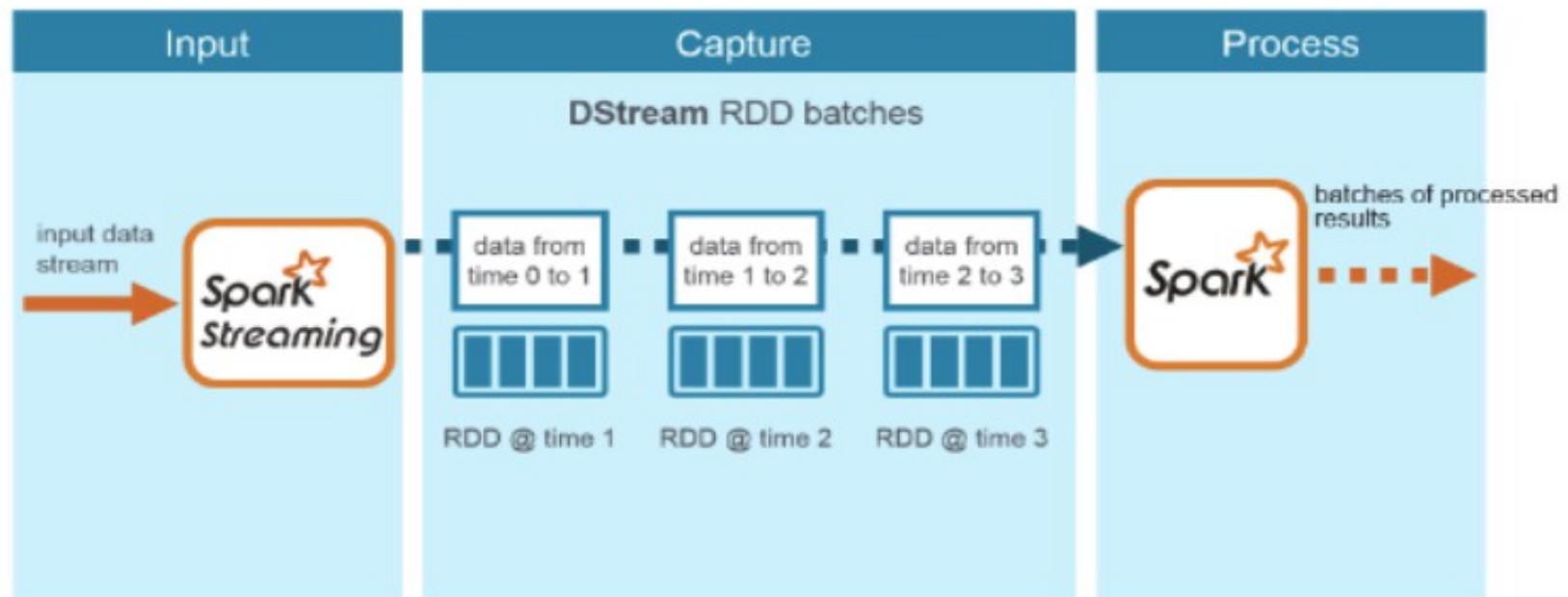


Spark Streaming Architecture

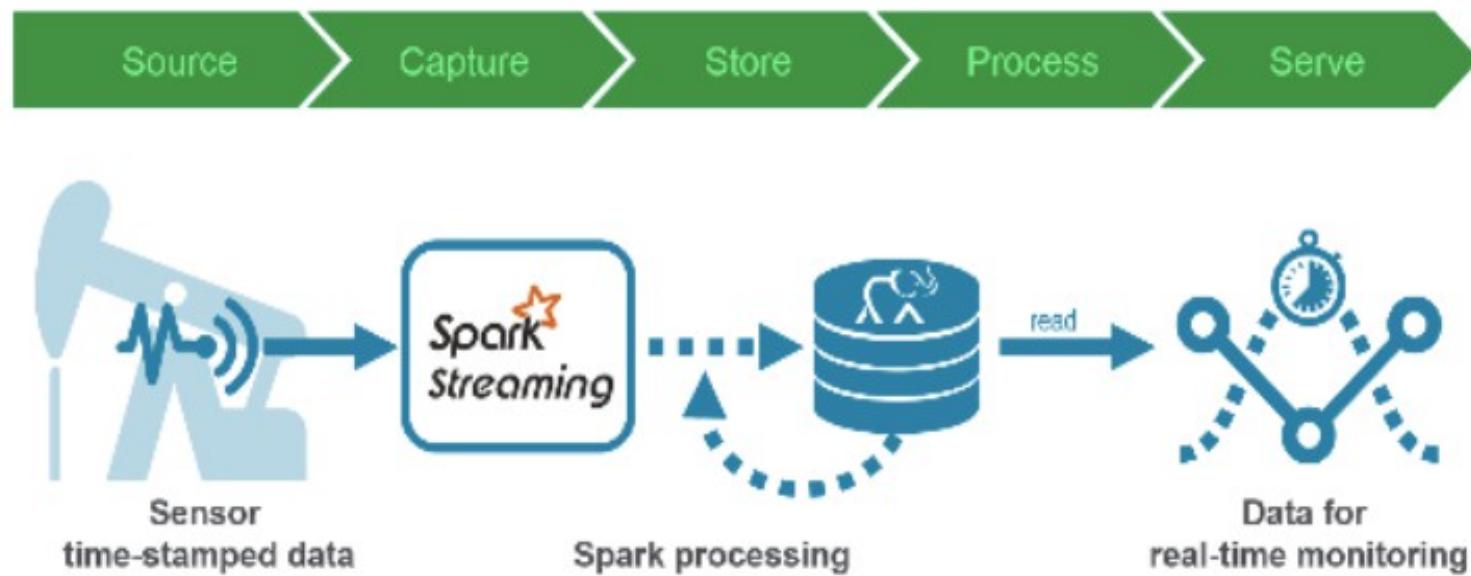


Processing Spark DStreams

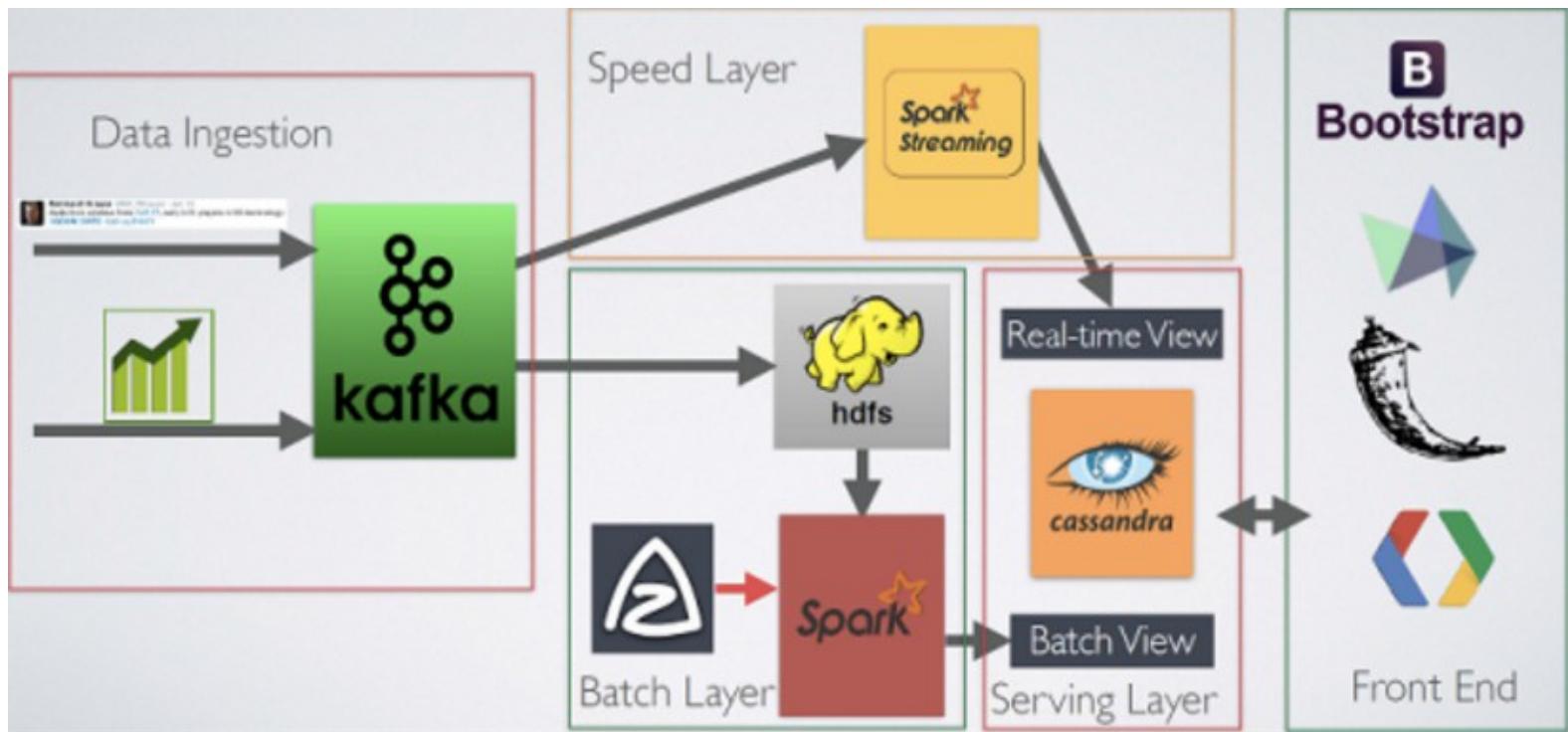
Processed results are pushed out in batches



Use Case: Time Series Data

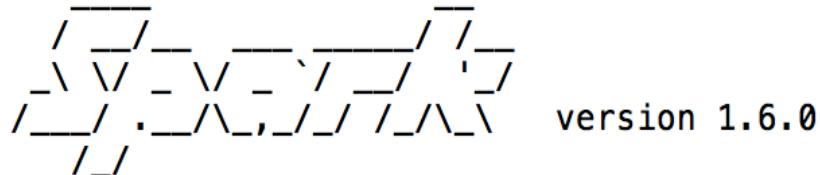


Use Case



Start Spark-shell with extra memory

```
[root@quickstart ~]# spark-shell --driver-memory 1G
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
Welcome to
```



version 1.6.0

WordCount using Spark Streaming

```
$ scala> :paste
import org.apache.spark.SparkConf
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.storage.StorageLevel
import StorageLevel._
import org.apache.spark._
import org.apache.spark.streaming._
import org.apache.spark.streaming.StreamingContext._
val ssc = new StreamingContext(sc, Seconds(2))
val lines = ssc.socketTextStream("localhost", 8585, MEMORY_ONLY)
val wordsFlatMap = lines.flatMap(_.split(" "))
val wordsMap = wordsFlatMap.map( w => (w,1))
val wordCount = wordsMap.reduceByKey( (a,b) => (a+b))
wordCount.print
ssc.start
```

Running the netcat server on another window

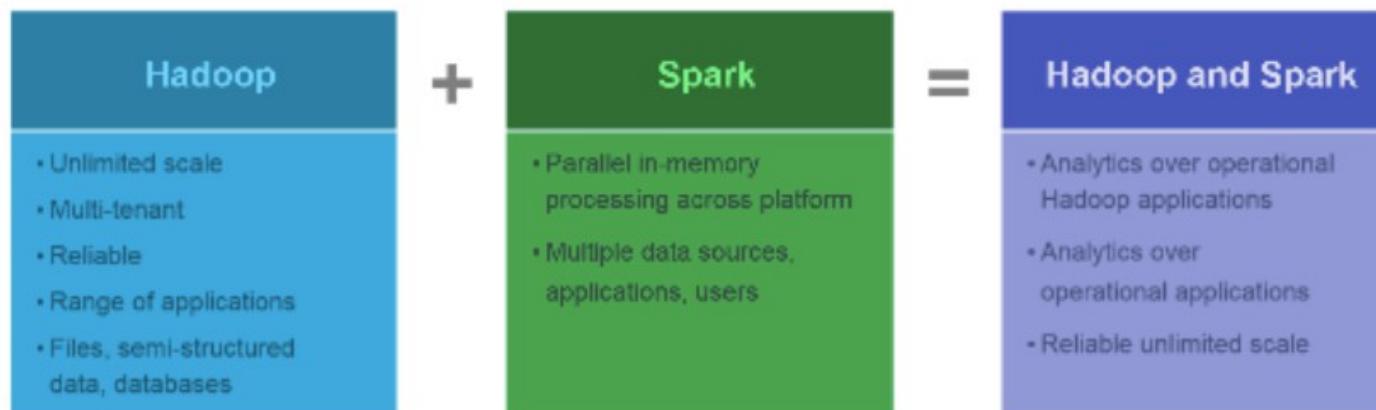
```
ubuntu@ip-172-31-30-238:~$ sudo su
root@ip-172-31-30-238:/home/ubuntu# docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED            STATUS              NAMES
581c45e85683        cloudera/quickstart:latest   "/usr/bin/docker-qui   About an hour ago
r ago   Up About an hour   0.0.0.0:8888->8888/tcp   backstabbing_lumiere
root@ip-172-31-30-238:/home/ubuntu# docker exec -i -t 581c45e85683 nc -lk 8585
```

```
test this
It is another test on Spark streaming. It is great
```

Time: 1465924608000 ms

(streaming.,1)
(Spark,1)
(great,1)
(is,2)
(test,1)
(another,1)
(on,1)
(It,2)

Hadoop + Spark



Challenge: Dataset

NYC's Taxi Trip Data

<http://www.andresmh.com/nyctaxitrips/>

NYC Taxi Trips

Data obtained through a FOIA request

[View the Project on GitHub](#)
andresmh/nyctaxitrips

[View On
GitHub](#)

[Read the
story](#)

[Check the
viz](#)

NYC Taxi Data

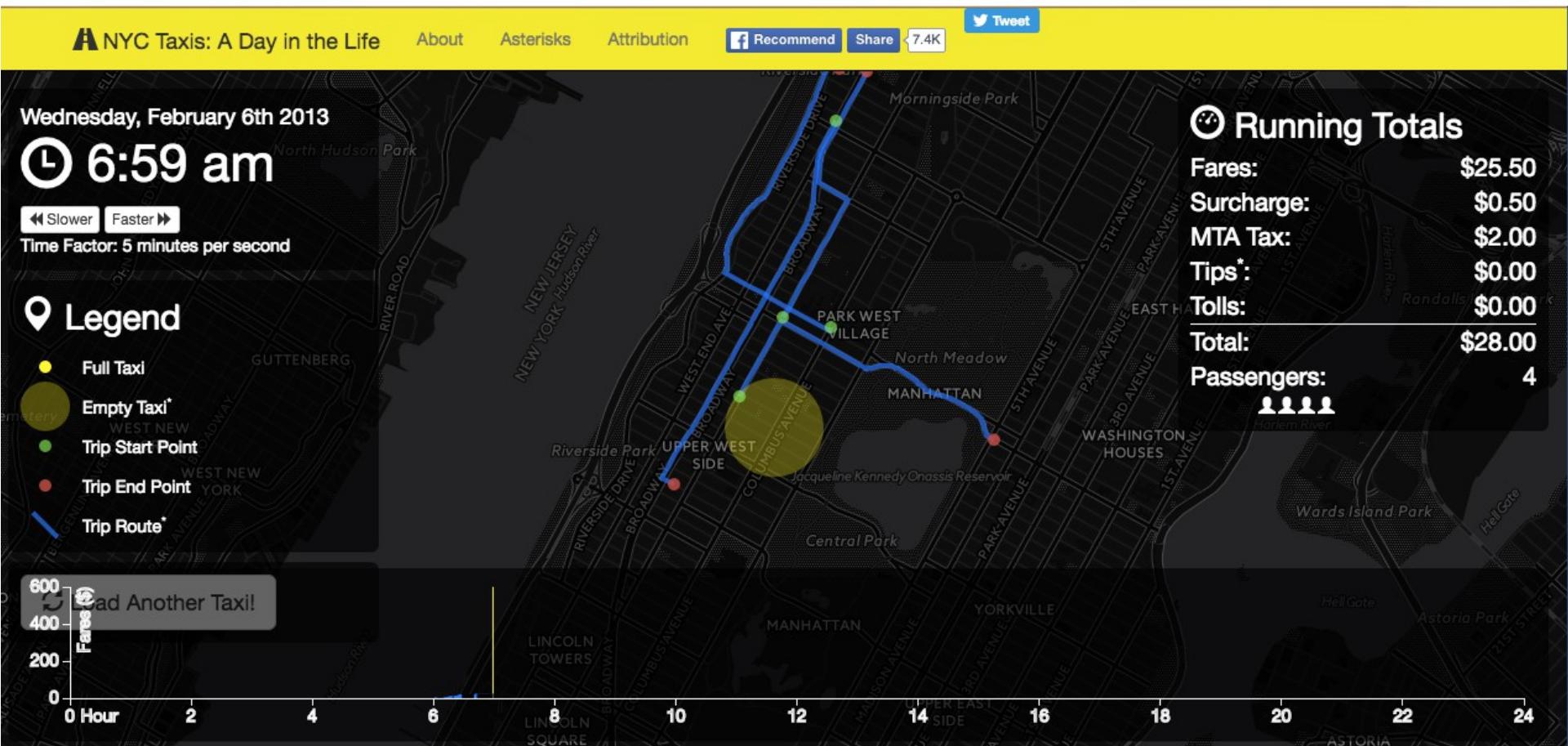
Trips

1. [trip_data_1.csv.zip](#)
2. [trip_data_2.csv.zip](#)
3. [trip_data_3.csv.zip](#)
4. [trip_data_4.csv.zip](#)
5. [trip_data_5.csv.zip](#)
6. [trip_data_6.csv.zip](#)
7. [trip_data_7.csv.zip](#)
8. [trip_data_8.csv.zip](#)
9. [trip_data_9.csv.zip](#)
10. [trip_data_10.csv.zip](#)
11. [trip_data_11.csv.zip](#)
12. [trip_data_12.csv.zip](#)

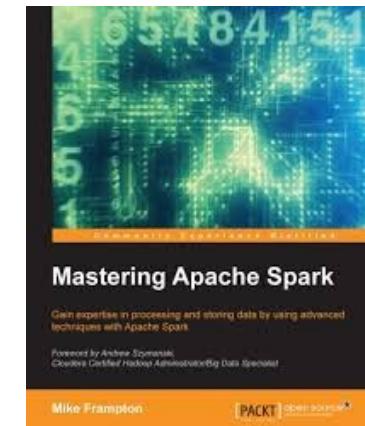
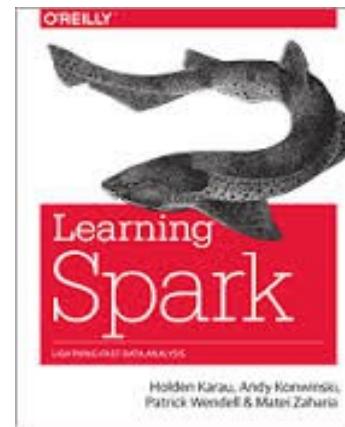
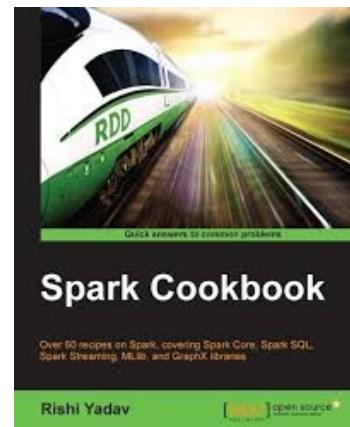
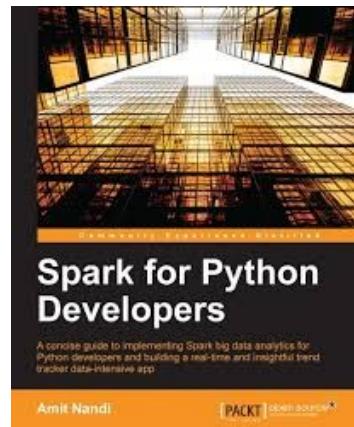


NYC Taxi : A Day in Life

<http://nyctaxi.herokuapp.com/>



Recommended Books





Big Data using Hadoop Workshop

27-28 July 2016

<http://www.imcinstitute.com/hadoop> Tel: 088-192-7975



Instructor:
Dr.Thanachart
Numnonda

Hadoop



HIVE



Hadoop



Fee: 5,500 Baht
(Early Bird)
Ex. VAT

Fee: include
Lunch Break
Course Material

Venue
Connection
@MRT Ladprao

Thank you

www.imcinstitute.com
www.facebook.com/imcinstitute