



Day 1

Introduction to Cloud Computing with Amazon EC2 and Apache Hadoop

Prof. Judy Qiu, Saliya Ekanayake, and Andrew Younge

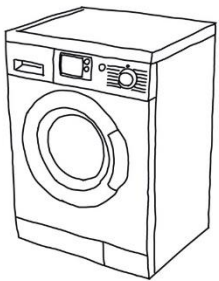
Presented By
Saliya Ekanayake



Cloud Computing

- What's Cloud?
 - Defining this is not worth the time
 - Ever heard of The Blind Men and The Elephant?
 - If you still need one, see NIST definition next slide
 - The idea is to consume X as-a-service, where X can be
 - Computing, storage, analytics, etc.
 - X can come from 3 categories
 - Infrastructure-as-a-S, Platform-as-a-Service, Software-as-a-Service

← Classic
Computing →



My washer
My bleach
I wash

IaaS



Rent a washer or two or three
My bleach
I wash

← Cloud
Computing →

PaaS



I tell,
comforter → dry clean
shirts → regular clean

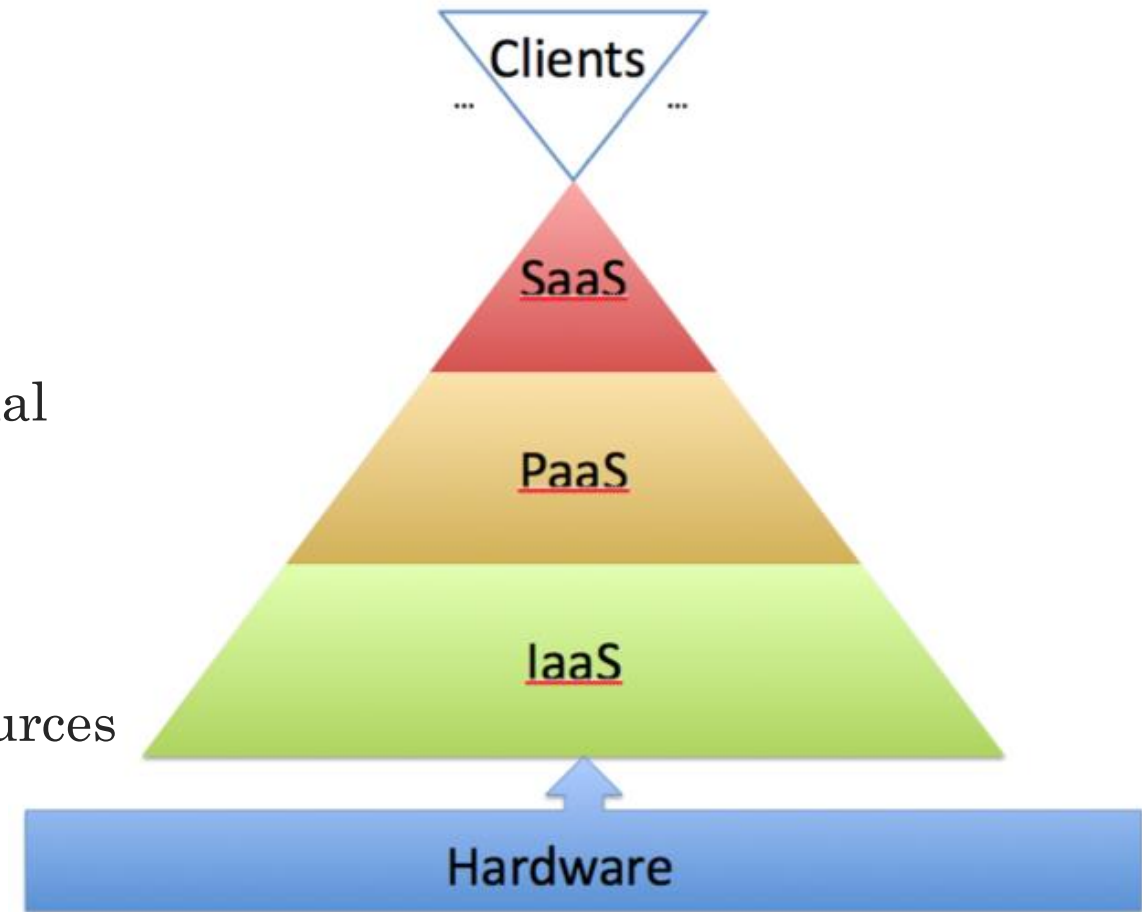
SaaS



Put my clothes in and
they magically appear
clean the next day

The Three Categories

- Software-as-a-Service
 - Provides web-enabled software
 - Ex: Google Gmail, Docs, etc
- Platform-as-a-Service
 - Provides scalable computing environments and runtimes for users to develop large computational and big data applications
 - Ex: Hadoop MapReduce
- Infrastructure-as-a-Service
 - Provide virtualized computing and storage resources in a dynamic, on-demand fashion.
 - Ex: Amazon Elastic Compute Cloud



The NIST Definition of Cloud Computing?

- *“Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”*
 - *On-demand self-service, broad network access, resource pooling, rapid elasticity, measured service,*
 - <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>
- However, formal definitions may not be very useful. We need hands on experience!



Cloud Computing

- Why Cloud?
 - Cost-effective
 - No upfront cost – pay as you go model
 - Elastic
 - On demand scaling
 - Maintenance free
 - Experienced people maintain it for you
 - Flexible
 - Mix and match architectures
 - Secure
 - Simple Programming Models and Services
 - Not always, but built-in support for many data analytic tasks



I Like Clouds. What Are My Options?

- Major Cloud Providers
 - Amazon <https://aws.amazon.com/>
 - Microsoft <https://azure.microsoft.com/en-us/>
 - Google <https://cloud.google.com/>
- Amazon vs. Microsoft vs. Google
 - <http://cloudacademy.com/blog/public-cloud-war-aws-vs-azure-vs-google/>
 - <https://www.youtube.com/watch?v=342KEaxFVjM>
- Other Providers
 - <http://cloud-computing.softwareinsider.com/>



Grants for Educators – Amazon

- Amazon AWS Educate <https://aws.amazon.com/education/awseducate/>

AWS Educate: Program Benefits

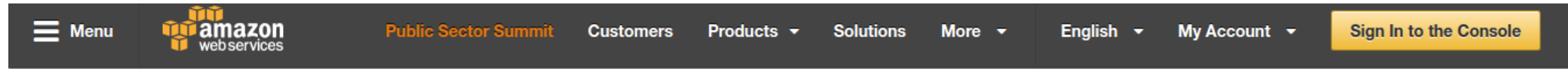
Access cloud content, training, collaboration tools, and AWS technology at no cost by joining AWS Educate today.

	Educators	Students
AWS Credits (annually renewable)	<ul style="list-style-type: none">• \$200 in AWS credits per educator – at member institutions• \$75 in AWS credits per educator – at non-member institutions	<ul style="list-style-type: none">• \$100 in AWS credits per student – at member institutions• \$35 in AWS credits per student – at non-member institutions
AWS Training	<ul style="list-style-type: none">• Free access to labs• Free AWS Technical Essentials eLearning course• 50% off instructor-led training provided by AWS in the United States, Brazil or Japan• 50% off AWS certification exams	<ul style="list-style-type: none">• Free access to labs
Curated Content	<ul style="list-style-type: none">• Free access to AWS content for classes• Free access to content contributed by leading educators	<ul style="list-style-type: none">• Free access to AWS content for homework, labs, or self-study
Collaboration Tools	<ul style="list-style-type: none">• Educator Collaboration Portal access• Virtual and in-person events• Contribute and rate content• Private and public discussion forums• Provide feedback on AWS Educate	<ul style="list-style-type: none">• Student Portal access• Virtual and in-person event to gather information, share best practices, and network• Provide feedback on AWS Educate



Grants for Educators – Amazon

- Amazon AWS Educate <http://aws.amazon.com/education/awseducate/apply/>



Apply for AWS Educate

It just takes a few minutes to apply for access to AWS Educate. Students, educators and administrators, just choose your category below and provide some basic information. It's that easy.



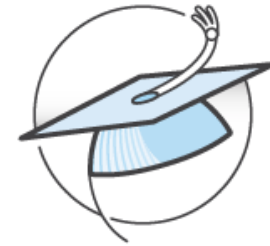
Educational Institutions

Apply for AWS Educate for Institutions



Educators

Apply for AWS Educate for Educators



Students

Apply for AWS Educate for Students

Have questions about signing up?

[Contact us today »](#)



IN
S

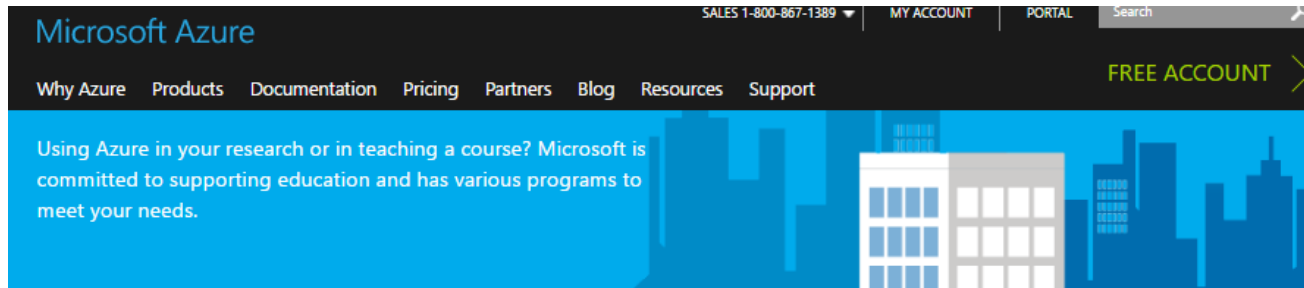
6/10/2016

Grants for Educators – Amazon

- Amazon offers credits to institutions, instructors, and students to use Amazon Web Services for free.
- Can apply for up to \$200 in instructor credits, \$100 in student credits if you are a member institution.
 - Must have class website with curriculum and members for verification
 - Apply with school .edu email address
- Applications processed in around 48 hours.
- Given a promotion code that's easily applied to your Amazon account.
- We are using AWS Educate credits for this workshop!



Grants for Educators – Microsoft



Educators

Empower faculty to leverage Microsoft Azure in teaching cutting edge courses



Virtual Machines: Run Windows or Linux virtual machines in the cloud



Mobile Services: Includes features that accelerate the development of mobile applications



Media Services: Create, manage or distribute media



Cloud Services: Build or extend existing enterprise applications



Big Data: Process enormous amounts of data

[See all services >](#)

The Educator Grant is a program designed specifically to provide access to Microsoft Azure to college and university professors teaching advanced courses. As part of the program, faculty teaching Azure in their curricula are awarded subscriptions to support their course.

To apply for an Educator Grant fill out this simple application form.

[Apply now >](#)

Join the over 750 universities in 74 countries that have participated in the program.

See all services at

<https://azure.microsoft.com/en-us/services/>

Apply at

<https://azure.microsoft.com/en-us/community/education/>


Hands-on 1

Getting Started with Amazon AWS



Amazon Web Services

<https://aws.amazon.com>

Menu  English My Account Create an AWS Account

Go to AWS

Create a new account or log-in to existing account

Gain free, hands-on experience with AWS for 12 months

Learn more about the AWS Free Tier »

Get Started with AWS for free

Create a Free Account

[View AWS Free Tier Details »](#)

Amazon Web Service

https://www.amazon.com/ap/signin?openid.assoc_handle



Sign In or Create an AWS Account

What is your email (phone for mobile accounts)?

E-mail or mobile number:

☐ I am a new user.

☒ I am a returning user and my password is:

[Sign in using our secure server](#)

[Forgot your password?](#)

Learn more about [AWS Identity and Access Management](#) and [AWS Multi-Factor Authentication](#) for additional security for your AWS Account. View full [AWS Free Usage T](#)



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

☐ Company Account ☒ Personal Account

* Required Fields

1

Full Name*

Country*


Address*

City*

State / Province or Region*

Postal Code*

Phone Number*

Security Check 

Please type the characters as shown above

AWS Customer Agreement

☐ Check here to indicate that you have read and agree to the terms of the [AWS Customer Agreement](#)

Payment Information

Please enter your payment information below. You will be able to try a broad set of AWS products for free via the Free Tier. We will only bill your credit or debit card for usage that is not covered by our Free Tier.

[Frequently Asked Questions](#)

AWS Free Tier	Compute Amazon EC2	Storage Amazon S3	Database Amazon RDS
free for 1 year	750hrs/month*	5GB	750hrs/month*

[*View full offer details](#)

Credit/Debit Card Number

Expiration Date

Cardholder's Name

- ☒ Use my contact address
 (
- ☐ Use a new address

Identity Verification

You will be called immediately by an automated system and prompted to enter the PIN number provided.

1. Provide a telephone number

Please enter your information below and click the "Call Me Now" button.

Country Code Phone Number Ext

2. Call in progress

3. Identity verification complete

Support Plan

4

AWS Support offers a selection of plans to meet your needs. All plans provide 24x7 access to customer service, documentation, whitepapers, and support forums. For access to technical support and additional help you plan, deploy, and optimize your AWS environment, we recommend selecting a support plan that aligns with your AWS usage.

Please Select One

- ☒ **Basic**
 Description: Customer Service for account and billing questions and access to the AWS Community Forums.
 Price: Included
- ☐ **Developer**
 Use case: Experimenting with AWS
 Description: One primary contact may ask technical questions through Support Center and get a response within 12–24 hours during local business hours.
 Price: \$49/month
- ☐ **Business**
 Use case: Production use of AWS
 Description: 24x7 support by phone and chat, 1-hour response to urgent support cases, and help with common third-party software. Full access to AWS Trusted Advisor for optimizing your AWS infrastructure, and access to the AWS Support API for automating your support cases and retrieving Trusted Advisor results.
 Price: \$100/month
- ☐ **Enterprise**
 Use case: Mission-critical use of AWS
 Description: All the features of the Business support plan, plus an assigned Technical Account Manager (TAM) who provides proactive guidance and best practices to help plan, develop, and run your AWS solutions, a Support Concierge who provides billing and account analysis and assistance, access to Infrastructure Event Management to support product launches, seasonal promotions/events, and migrations, and 15-minute response to critical support cases with prioritized case handling.
 Price: \$15,000/month
 If you select this option, customer support will contact you within 48 hours to discuss your needs and finalize the sign-up. Support resources will be available when sign-up is finalized, and no charges will be incurred until that time.
- To explore all features and benefits of AWS Support, including plan comparisons and pricing samples, [click here](#).

5

Welcome to Amazon Web Services

Thank you for creating an Amazon Web Services Account. We are activating your account, which should only takes a few minutes. You will receive an email when this is complete.

Get Started with AWS Technical Documentation



Launch a Linux Virtual Machine



Store Your Files in the Cloud



Launch a WordPress Website



Launch a Web Application

[View all documentation >>](#)



INDIANA UNIVERSITY
SCHOOL

6/10/2016

ROUTING



Quick Starts



Build a web app



Launch a Virtual Machine
(EC2 Instance)



Back up your files
[Learn More](#)



Build a backend for your
mobile app



Host a static website



Analyze big data
[Learn More](#)

If all goes well, you
should be able to see
this page

AWS Services SHOW CATEGORIES

COMPUTE

[EC2](#)
[EC2 Container Service](#)
[Elastic Beanstalk](#)
[Lambda](#)

STORAGE & CONTENT DELIVERY

[S3](#)
[CloudFront](#)
[Elastic File System](#) PREVIEW
[Glacier](#)
[Snowball](#)
[Storage Gateway](#)

DATABASE

[RDS](#)
[DynamoDB](#)
[ElastiCache](#)
[Redshift](#)
[DMS](#)

DEVELOPER TOOLS

[CodeCommit](#)
[CodeDeploy](#)
[CodePipeline](#)

MANAGEMENT TOOLS

[CloudWatch](#)
[CloudFormation](#)
[CloudTrail](#)
[Config](#)
[OpsWorks](#)
[Service Catalog](#)
[Trusted Advisor](#)

SECURITY & IDENTITY

[IAM](#)
[Directory Service](#)
[Inspector](#)
[WAF](#)
[Certificate Manager](#)

INTERNET OF THINGS

[AWS IoT](#)

GAME DEVELOPMENT

[GameLift](#)

MOBILE SERVICES

[Mobile Hub](#)
[Cognito](#)
[Device Farm](#)
[Mobile Analytics](#)
[SNS](#)

APPLICATION SERVICES

[API Gateway](#)
[AppStream](#)
[CloudSearch](#)
[Elastic Transcoder](#)
[SES](#)
[SQS](#)
[SWF](#)

GETTING STARTED

Read our [documentation](#) or view our [training](#) to learn more about AWS.

AWS CONSOLE MOBILE APP

View your resources on the go with our AWS Console mobile app, available from [Amazon Appstore](#), [Google Play](#), or [iTunes](#).

AWS MARKETPLACE

Find and buy software, launch with 1-Click and pay by the hour.

FEEDBACK

Let us know what you think about new Console Home.

Service Health

[View Dashboard](#)



All services are operating normally.
Updated Jun 08 2016 14:36:00 GMT-0400



Hands-on 1

Questions?



Amazon Web Services





- Grew out of Amazon's need to rapidly provision and configure machines of standard configurations for its own business.
- Early 2000s – Both private and shared data centers began using virtualization to perform “server consolidation”
- 2003 – Internal memo by Chris Pinkham describing an “infrastructure service for the world.”
- 2006 – S3 first deployed in the spring, EC2 in the fall
- 2008 – Elastic Block Store available.
- 2009 – Relational Database Service
- 2012 – DynamoDB
- 2015 – Amazon ECS









AWS Services

Amazon Web Services





Compute

-  **EC2**
Virtual Servers in the Cloud
-  **EC2 Container Service**
Run and Manage Docker Containers
-  **Elastic Beanstalk**
Run and Manage Web Apps
-  **Lambda**
Run Code in Response to Events

Storage & Content Delivery

-  **S3**
Scalable Storage in the Cloud
-  **CloudFront**
Global Content Delivery Network
-  **Elastic File System** PREVIEW
Fully Managed File System for EC2
-  **Glacier**
Archive Storage in the Cloud
-  **Snowball**
Large Scale Data Transport
-  **Storage Gateway**
Hybrid Storage Integration








Database

-  **RDS**
Managed Relational Database Service
-  **DynamoDB**
Managed NoSQL Database
-  **ElastiCache**
In-Memory Cache
-  **Redshift**





Developer Tools

-  **CodeCommit**
Store Code in Private Git Repositories
-  **CodeDeploy**
Automate Code Deployments
-  **CodePipeline**
Release Software using Continuous Delivery


Management Tools

-  **CloudWatch**
Monitor Resources and Applications
-  **CloudFormation**
Create and Manage Resources with Templates
-  **CloudTrail**
Track User Activity and API Usage
-  **Config**
Track Resource Inventory and Changes
-  **OpsWorks**
Automate Operations with Chef
-  **Service Catalog**
Create and Use Standardized Products
-  **Trusted Advisor**
Optimize Performance and Security


Security & Identity

-  **Identity & Access Management**
Manage User Access and Encryption Keys
-  **Directory Service**
Host and Manage Active Directory
-  **Inspector**
Analyze Application Security
-  **WAF**






Internet of Things

-  **AWS IoT**
Connect Devices to the Cloud







Game Development

-  **GameLift**
Deploy and Scale Session-based Multiplayer Games

Mobile Services

-  **Mobile Hub**
Build, Test, and Monitor Mobile Apps
-  **Cognito**
User Identity and App Data Synchronization
-  **Device Farm**
Test Android, iOS, and Web Apps on Real Devices in the Cloud
-  **Mobile Analytics**
Collect, View and Export App Analytics
-  **SNS**
Push Notification Service

Application Services

-  **API Gateway**
Build, Deploy and Manage APIs
-  **AppStream**
Low Latency Application Streaming
-  **CloudSearch**
Managed Search Service
-  **Elastic Transcoder**
Easy-to-Use Scalable Media Transcoding
-  **SES**
Email Sending and Receiving Service
-  **SQS**
Message Queue Service



Get Certified!

- <https://aws.amazon.com/certification/>

AWS Certification

AWS Certifications recognize IT professionals with the technical skills and expertise to design, deploy, and operate applications and infrastructure on AWS. Exams are offered in multiple languages at testing centers around the world.



Why Get Certified?

Show You're an Expert

AWS Certification is an industry-recognized credential that shows you have the expertise to design, deploy, or operate applications and infrastructure on AWS.

Advance Your Career

AWS Certification shows customers, employers, and recruiters that you have the skills and knowledge to build AWS solutions.

Join the Community

Put the AWS Certified logo on your LinkedIn profile, email, and website—become a member of our LinkedIn AWS Certified community.

Roadmap

Associate Exams

Professional Exams

Earn your Associate level certification and then advance to Professional level within a given role.



Solutions Architect

AWS Certified Solutions
Architect - Associate

AWS Certified Solutions
Architect - Professional



Developer

AWS Certified
Developer - Associate

AWS Certified DevOps Engineer - Professional



SysOps Administrator

AWS Certified SysOps
Administrator - Associate



INDIANA UNIVERSITY BLOOMINGTON
SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

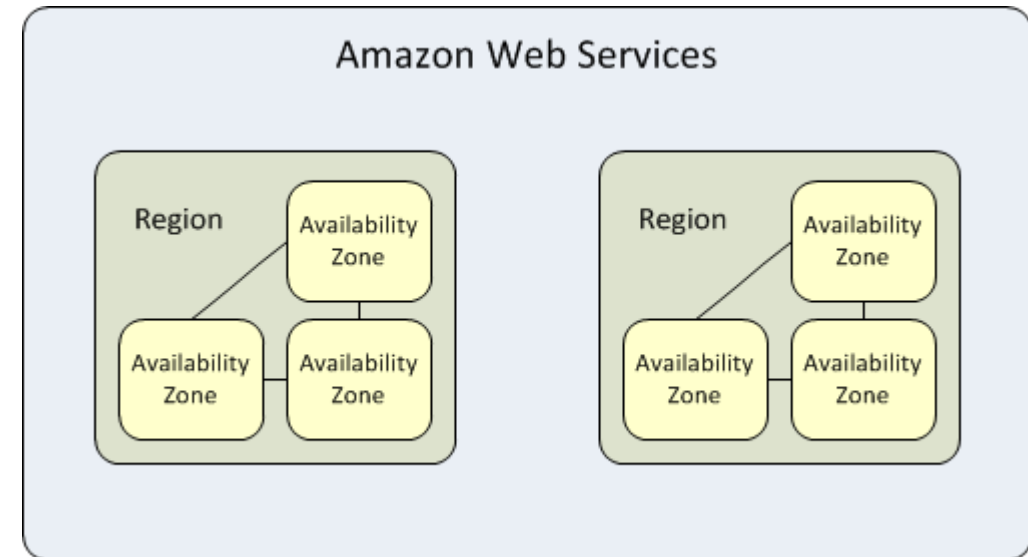
Amazon Elastic Compute Cloud (EC2)

- Amazon EC2 is a central component of the Amazon Web Services
- Provides virtualized computing resources on-demand.
- Creates and manages VM instances, thereby renting computing services based on resource requests
- Interaction with other AWS services such as S3, EBS, etc.
- Public Infrastructure-as-a-Service



Terminology

- Instance
 - One running virtual machine.
- Instance Type
 - hardware configuration: cores, memory, disk.
- Instance Store Volume
 - Temporary disk associated with instance.
- Image (AMI)
 - Stored bits which can be turned into instances.
- Key Pair
 - Credentials used to access VM from command line.
- Region
 - Geographic location, price, laws, network locality.
- Availability Zone
 - Subdivision of region the is fault-independent.
 - <http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html>



Model	vCPU	CPU Credits / hour	Mem (GiB)	Storage (GB)
t2.micro	1	6	1	EBS Only
t2.small	1	12	2	EBS Only
t2.medium	2	24	4	EBS Only
c3.large		2	3.75	2 x 16
c3.xlarge		4	7.5	2 x 40
c3.2xlarge		8	15	2 x 80
c3.4xlarge		16	30	2 x 160
c3.8xlarge		32	60	2 x 320

Use Cases

High performance front-end fleets, web-servers, on-demand batch processing, distributed analytics, high performance science and engineering applications, ad serving, batch processing, MMO gaming, video encoding, and distributed analytics.

Model	vCPU	Mem (GiB)	SSD Storage (GB)
m3.medium	1	3.75	1 x 4
m3.large	2	7.5	1 x 32
m3.xlarge	4	15	2 x 40
m3.2xlarge	8	30	2 x 80
r3.large	2	15.25	1 x 32
r3.xlarge	4	30.5	1 x 80
r3.2xlarge	8	61	1 x 160
r3.4xlarge	16	122	1 x 320
r3.8xlarge	32	244	2 x 320

Use Cases

We recommend memory-optimized instances for high performance databases, distributed memory caches, in-memory analytics, genome assembly and analysis, larger deployments of SAP, Microsoft SharePoint, and other enterprise applications.



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

EC2 Pricing Model

- Free Usage Tier
- On-Demand Instances
 - Start and stop instances whenever you like, costs are rounded up to the nearest hour. (Worst price)
- Reserved Instances
 - Pay up front for one/three years in advance. (Best price)
 - Unused instances can be sold on a secondary market.
- Spot Instances
 - Specify the price you are willing to pay, and instances get started and stopped without any warning as the market changes. (Kind of like Condor!)

<http://aws.amazon.com/ec2/pricing/>



Free Usage Tier

- 750 hours of EC2 running Linux, RHEL, or SLES t2.micro instance usage
- 750 hours of EC2 running Microsoft Windows Server t2.micro instance usage
- 750 hours of Elastic Load Balancing plus 15 GB data processing
- 30 GB of Amazon Elastic Block Storage in any combination of General Purpose (SSD) or Magnetic, plus 2 million I/Os (with Magnetic) and 1 GB of snapshot storage
- 15 GB of bandwidth out aggregated across all AWS services
- 1 GB of Regional Data Transfer



Surprisingly, you can't scale up that large.

Q: How many instances can I run in Amazon EC2?

You are limited to running up to 20 On-Demand Instances, purchasing 20 Reserved Instances, and requesting 5 Spot Instances per region. New AWS accounts may start with limits that are lower than the limits described here. Certain instance types are further limited per region as follows:

Instance Type	On-Demand Limit	Reserved Limit	Spot Limit
cg1.4xlarge	2	20	5
hi1.4xlarge	2	20	5
hs1.8xlarge	2	20	Not offered
cr1.8xlarge	2	20	5
g2.2xlarge	5	20	5
r3.4xlarge	10	20	5



Simple Storage Service (S3)

- A **bucket** is a container for objects and describes location, logging, accounting, and access control. A bucket can hold any number of **objects**, which are files of up to 5TB. A bucket has a name that must be **globally unique**.
- Fundamental operations corresponding to HTTP actions:
 - `http://bucket.s3.amazonaws.com/object`
 - POST a new object or update an existing object.
 - GET an existing object from a bucket.
 - DELETE an object from the bucket
 - LIST keys present in a bucket, with a filter.
- A bucket has a **flat directory structure** (despite the appearance given by the interactive web interface.)



Bucket Properties

- Versioning – If enabled, POST/DELETE result in the creation of new versions without destroying the old.
- Lifecycle – Delete or archive objects in a bucket a certain time after creation or last access or number of versions.
- Access Policy – Control **when and where** objects can be accessed.
- Access Control – Control who **may** access objects in this bucket.
- Logging – Keep track of how objects are accessed.
- Notification – Be notified when failures occur.



S3 Weak Consistency Model

From Amazon developer API:

- “Updates to a single key are **atomic....**”
- Amazon S3 achieves high availability by replicating data across multiple servers within Amazon's data centers.
- If a PUT request is successful, your data is safely stored.
 - However, information about the changes must replicate across Amazon S3, which can take some time

	Standard Storage	Reduced Redundancy Storage	Glacier Storage
First 1 TB / month	\$0.0300 per GB	\$0.0240 per GB	\$0.0100 per GB
Next 49 TB / month	\$0.0295 per GB	\$0.0236 per GB	\$0.0100 per GB
Next 450 TB / month	\$0.0290 per GB	\$0.0232 per GB	\$0.0100 per GB
Next 500 TB / month	\$0.0285 per GB	\$0.0228 per GB	\$0.0100 per GB
Next 4000 TB / month	\$0.0280 per GB	\$0.0224 per GB	\$0.0100 per GB
Over 5000 TB / month	\$0.0275 per GB	\$0.0220 per GB	\$0.0100 per GB



Elastic Block Store

- An EBS volume is a **virtual disk** of a fixed size with a block read/write interface. It can be **mounted** as a filesystem on a running EC2 instance where it can be **updated incrementally**. Unlike an instance store, an EBS volume is **persistent**.
- (Compare to an S3 object, which is essentially a file that must be accessed in its entirety.)
- Fundamental operations:
 - CREATE a new volume (1GB-1TB)
 - COPY a volume from an existing EBS volume or S3 object.
 - MOUNT on one instance at a time.
 - SNAPSHOT current state to an S3 object.

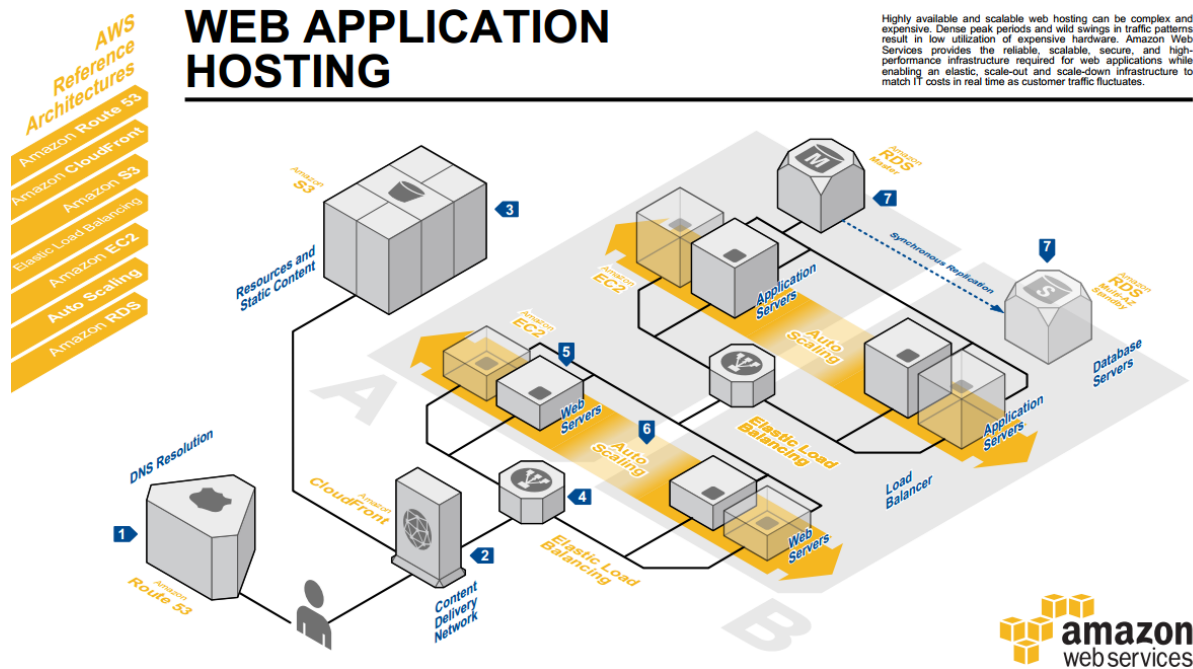


Amazon EBS Volume Types

Volume Type	EBS General Purpose (SSD)	EBS Provisioned IOPS (SSD)	EBS Magnetic
Use Cases	Boot volumes Small to Med DBs Dev and Test	I/O intensive Relational DBs NoSQL DBs	Infrequent Data Access
Storage Media	SSD-backed	SSD-backed	Magnetic disk-backed
Max Volume Size	1TB	1TB	1TB
Max IOPS/volume	3,000 (burst)	4,000	40 - 200
Max throughput/volume	128MBps	128MBps	40 - 90MBps
Max IOPS/instance	48,000	48,000	48,000
Max throughput/instance	800MBps	800MBps	800MBps
API Name	gp2	io1	standard
Price*	\$.10/GB - Month	\$.125/GB - Month \$.065/provisioned IOPS	\$.05/GB - Month \$.05/million I/O

Where to Find More Info?

- The Getting Started Guide
 - <http://docs.aws.amazon.com/gettingstarted/latest/awsgsg-intro/gsg-aws-intro.html>
- AWS Architecture Center
 - <https://aws.amazon.com/architecture/>



Hands-on 2

Launching EC2 Instances



Go to AWS.Amazon.com

Menu



AWS Summit: Santa Clara

Products

Solutions

Pricing

More ▾

English ▾

My Account ▾

Sign In to the Console

PRODUCTS & SERVICES

Amazon EC2 >

Product Details >

Instances >

Pricing >

Purchasing Options >

Developer Resources >

FAQs >

Getting Started >

Amazon EC2 Run Command >

RELATED LINKS

Amazon EC2 Dedicated Hosts

Amazon EC2 Spot Instances

Amazon EC2 Reserved Instances

Amazon EC2 Dedicated Instances

Amazon EC2 - Virtual Server Hosting

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale [cloud computing](#) easier for developers.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Introduction to Amazon EC2 (4:01)

Manage Your AWS Resources

Enter EC2 Console



AWS MONTHLY WEBINAR SERIES

Learn more about AWS products and services with AWS experts. Choose from a range of topics from introductory level to technical deep dives.

[Browse the Topics and](#)

Adobe Test & Target

Amazon Associates

Marketo

Omniture (Adobe Analytics)



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

Log into EC2 Dashboard

The screenshot shows the AWS Management Console interface for the EC2 Dashboard. The top navigation bar includes the AWS logo, 'Services', 'Edit', and user information (Andrew J. Younge, Oregon, Support). The left sidebar lists navigation options: EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES (Instances, Spot Requests, Reserved Instances, Scheduled Instances, Dedicated Hosts), IMAGES (AMIs, Bundle Tasks), ELASTIC BLOCK STORE (Volumes, Snapshots), and NETWORK & SECURITY (Security Groups, Elastic IPs, Placement Groups, Key Pairs, Network Interfaces). The main content area is titled 'Resources' and shows the following counts for the US West (Oregon) region: 1 Running Instances, 0 Elastic IPs, 0 Dedicated Hosts, 1 Snapshots, 1 Volumes, 0 Load Balancers, 1 Key Pairs, 6 Security Groups, and 0 Placement Groups. A blue box with a close icon contains the text: 'Build and run distributed, fault-tolerant applications in the cloud with Amazon Simple Workflow Service.' Below this is the 'Create Instance' section, which includes the text 'To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.' and a blue 'Launch Instance' button, which is circled in red. A note below states: 'Note: Your instances will launch in the US West (Oregon) region.' The bottom section shows 'Service Health' for 'US West (Oregon):' with a green checkmark and the text 'This service is operating normally'. The right sidebar contains 'Account Attributes' (Supported Platforms, VPC, Default VPC, vpc-bdc248d8, Resource ID length management), 'Additional Information' (Getting Started Guide, Documentation, All EC2 Resources, Forums, Pricing, Contact Us), and 'AWS Marketplace' (Find free software trial products in the AWS Marketplace from the EC2 Launch Wizard. Or try these popular AMIs: Tableau Server (10 users), Provided by Tableau). The bottom footer includes 'Feedback', 'English', copyright information (© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved.), 'Privacy Policy', and 'Terms of Use'. A 'Google Analytics' badge is also visible.




INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

Launch your first EC2 instance!

 **AWS** ▾ **Services** ▾ **Edit** ▾ Andrew J. Younge ▾ Oregon ▾ Support ▾

1. Choose AMI | 2. Choose Instance Type | 3. Configure Instance | 4. Add Storage | 5. Tag Instance | 6. Configure Security Group | 7. Review

Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

[Cancel and Exit](#)


Quick Start

My AMIs

AWS Marketplace


Community AMIs

☐ Free tier only ⓘ

**Amazon Linux**
Free tier eligible


Amazon Linux AMI 2016.03.1 (HVM), SSD Volume Type - ami-d0f506b0
The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages.
Root device type: ebs Virtualization type: hvm

Select
64-bit

**Red Hat**
Free tier eligible

Red Hat Enterprise Linux 7.2 (HVM), SSD Volume Type - ami-775e4f16
Red Hat Enterprise Linux version 7.2 (HVM), EBS General Purpose (SSD) Volume Type
Root device type: ebs Virtualization type: hvm

Select
64-bit

**SUSE Linux**
Free tier eligible

SUSE Linux Enterprise Server 12 SP1 (HVM), SSD Volume Type - ami-d2627db3
SUSE Linux Enterprise Server 12 Service Pack 1 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled.
Root device type: ebs Virtualization type: hvm

Select
64-bit

Select an Instance Type

AWS

Services

Edit

Andrew J. Young

Oregon

Support

1. Choose AMI

2. Choose Instance Type

3. Configure Instance

4. Add Storage

5. Tag Instance

6. Configure Security Group

7. Review

Step 2: Choose an Instance Type

Filter by:

All instance types

Current generation

Show/Hide Columns

Currently selected: t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
<input type="checkbox"/>	General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate
<input checked="" type="checkbox"/>	General purpose	t2.micro Free tier eligible	1	1	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.large	2	8	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	m4.large	2	8	EBS only	Yes	Moderate
<input type="checkbox"/>	General purpose	m4.xlarge	4	16	EBS only	Yes	High

Cancel

Previous

Review and Launch

Next: Configure Instance Details

Feedback


English

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Privacy Policy

Terms of Use

Review your Instance settings, and Launch!

AWS

Services

Edit

Andrew J. YoungeOregonSupport

1. Choose AMI2. Choose Instance Type3. Configure Instance4. Add Storage5. Tag Instance6. Configure Security Group7. Review

Step 7: Review Instance Launch

▼ AMI Details

Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9abea4fb

Free tier eligible

Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).
Root Device Type: ebsVirtualization type: hvm

Edit AMI

▼ Instance Type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.micro	Variable	1	1	EBS only	-	Low to Moderate

Edit instance type

▶ Security Groups

Edit security groups

▶ Instance Details

Edit instance details

▶ Storage

Edit storage

▶ Tags

Edit tags

CancelPreviousLaunch

FeedbackEnglish

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy PolicyTerms of Use

Amazon uses SSH keypairs

- Amazon EC2 uses SSH keypairs to control access to VMs
- Consists of public key (known) and private key (secret)
- You select which public key to use, and log in with your private key.
- Can use many different keypairs

Select an existing key pair or create a new key pair ×

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Choose an existing key pair

Select a key pair

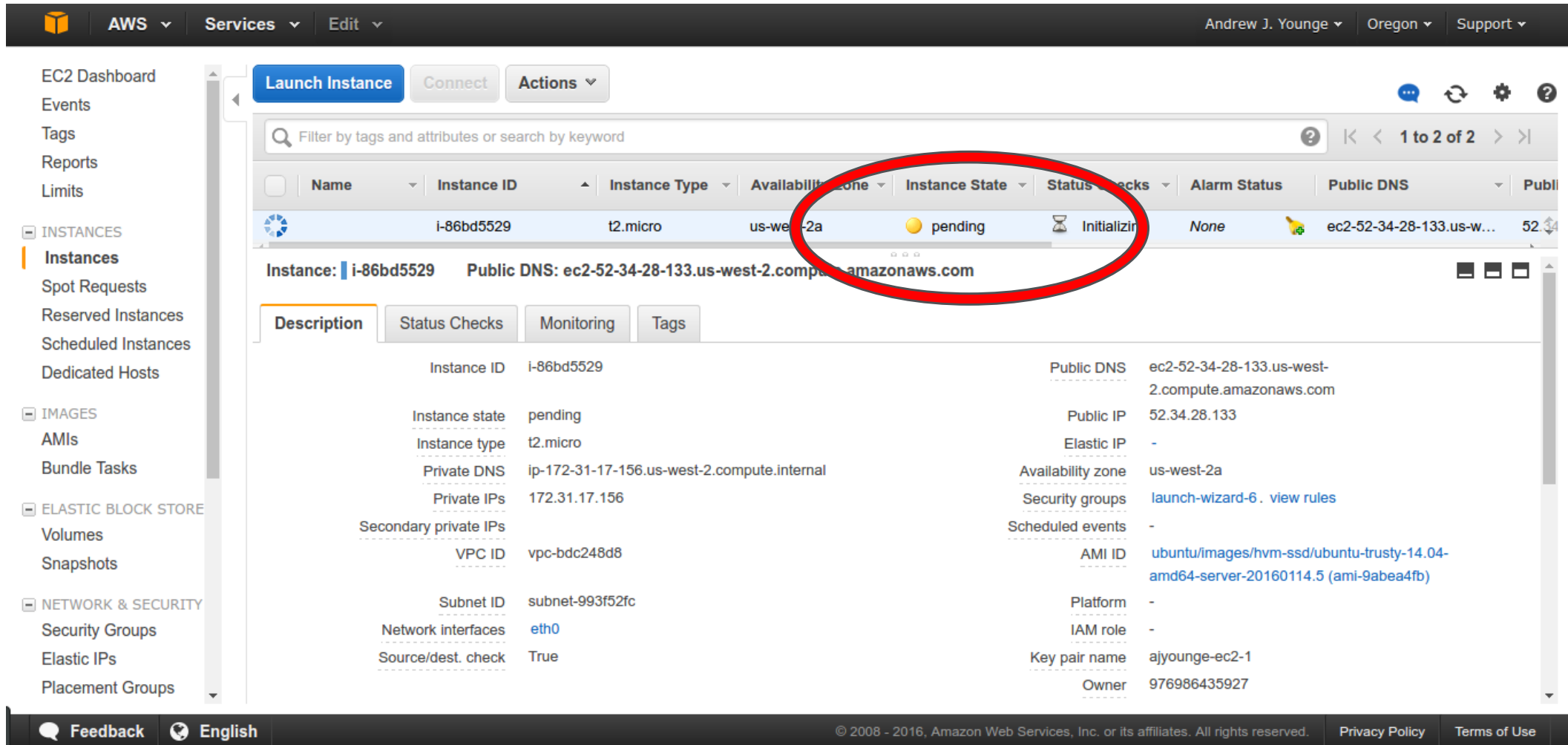
ajyounge-ec2-1

☒ I acknowledge that I have access to the selected private key file (ajyounge-ec2-1.pem), and that without this file, I won't be able to log into my instance.

Cancel Launch Instances



Booting your Instance...



The screenshot shows the AWS Management Console interface for the EC2 service. The top navigation bar includes the AWS logo, 'Services', 'Edit', and user information (Andrew J. Younge, Oregon, Support). The left sidebar lists navigation options: EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, Images, ELASTIC BLOCK STORE, and NETWORK & SECURITY. The main content area displays a table of EC2 instances. The first instance, 'i-86bd5529', is highlighted. Its 'Instance State' is 'pending', and its 'Status Checks' are 'Initializing'. A red circle is drawn around the 'pending' status and the 'Initializing' progress indicator. Below the table, the 'Description' tab is selected, showing details for the instance, including its ID, state, type, DNS, IPs, VPC, subnet, interfaces, and source/destination check.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
	i-86bd5529	t2.micro	us-west-2a	pending	Initializing	None	ec2-52-34-28-133.us-west-2.compute.amazonaws.com	52.34.28.133

Instance: i-86bd5529 Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com

Description Status Checks Monitoring Tags

Instance ID	i-86bd5529	Public DNS	ec2-52-34-28-133.us-west-2.compute.amazonaws.com
Instance state	pending	Public IP	52.34.28.133
Instance type	t2.micro	Elastic IP	-
Private DNS	ip-172-31-17-156.us-west-2.compute.internal	Availability zone	us-west-2a
Private IPs	172.31.17.156	Security groups	launch-wizard-6 . view rules
Secondary private IPs	-	Scheduled events	-
VPC ID	vpc-bdc248d8	AMI ID	ubuntu/images/hvm-ssd/ubuntu-trusty-14.04-amd64-server-20160114.5 (ami-9abaa4fb)
Subnet ID	subnet-993f52fc	Platform	-
Network interfaces	eth0	IAM role	-
Source/dest. check	True	Key pair name	ajyounge-ec2-1
		Owner	976986435927

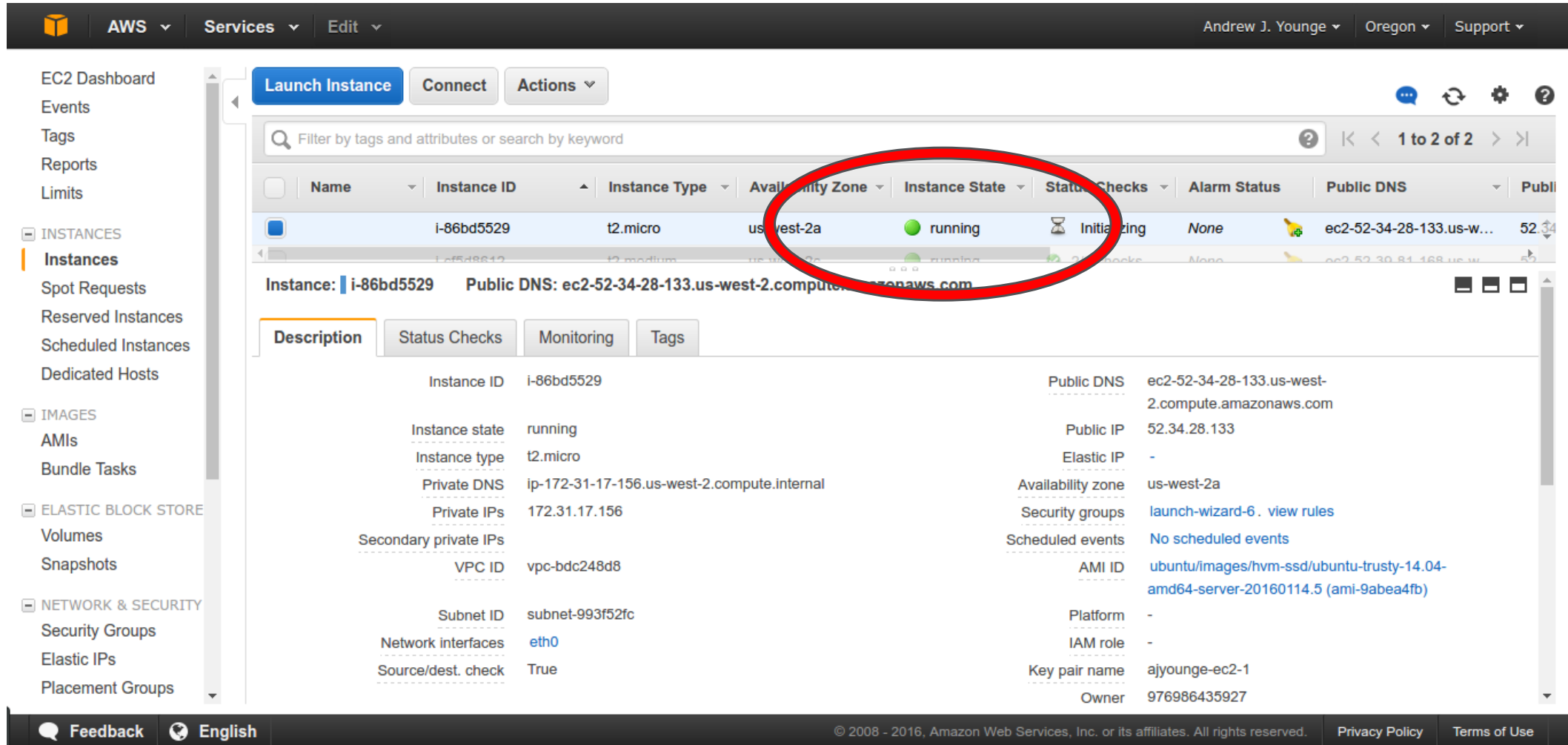


INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

Instance is running!



The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, 'AWS', 'Services', 'Edit', and user information 'Andrew J. Younge', 'Oregon', and 'Support'. The left sidebar contains navigation links for 'EC2 Dashboard', 'Events', 'Tags', 'Reports', 'Limits', 'INSTANCES', 'Instances', 'Spot Requests', 'Reserved Instances', 'Scheduled Instances', 'Dedicated Hosts', 'IMAGES', 'AMIs', 'Bundle Tasks', 'ELASTIC BLOCK STORE', 'Volumes', 'Snapshots', 'NETWORK & SECURITY', 'Security Groups', 'Elastic IPs', and 'Placement Groups'. The main content area displays a table of EC2 instances. The first instance, 'i-86bd5529', is highlighted in blue. A red circle is drawn around the 'Instance State' column for this instance, which shows a green dot and the word 'running'. Below the table, the 'Instance: i-86bd5529' details are shown, including the 'Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com'. The details are organized into tabs: 'Description', 'Status Checks', 'Monitoring', and 'Tags'. The 'Description' tab is active, showing various instance attributes.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
	i-86bd5529	t2.micro	us-west-2a	running	Initializing	None	ec2-52-34-28-133.us-w...	52.34.28.133

Instance: i-86bd5529 Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com

Description		Status Checks		Monitoring		Tags	
Instance ID	i-86bd5529	Public DNS	ec2-52-34-28-133.us-west-2.compute.amazonaws.com	Instance state	running	Public IP	52.34.28.133
Instance type	t2.micro	Elastic IP	-	Instance type	t2.micro	Availability zone	us-west-2a
Private DNS	ip-172-31-17-156.us-west-2.compute.internal	Security groups	launch-wizard-6 . view rules	Private DNS	ip-172-31-17-156	Scheduled events	No scheduled events
Private IPs	172.31.17.156	AMI ID	ubuntu/images/hvm-ssd/ubuntu-trusty-14.04-amd64-server-20160114.5 (ami-9abea4fb)	Secondary private IPs	-	Platform	-
VPC ID	vpc-bdc248d8	IAM role	-	VPC ID	vpc-bdc248d8	Key pair name	ajyounge-ec2-1
Subnet ID	subnet-993f52fc	Owner	976986435927	Subnet ID	subnet-993f52fc		
Network interfaces	eth0			Network interfaces	eth0		
Source/dest. check	True			Source/dest. check	True		



INDIANA UNIVERSITY BLOOMINGTON

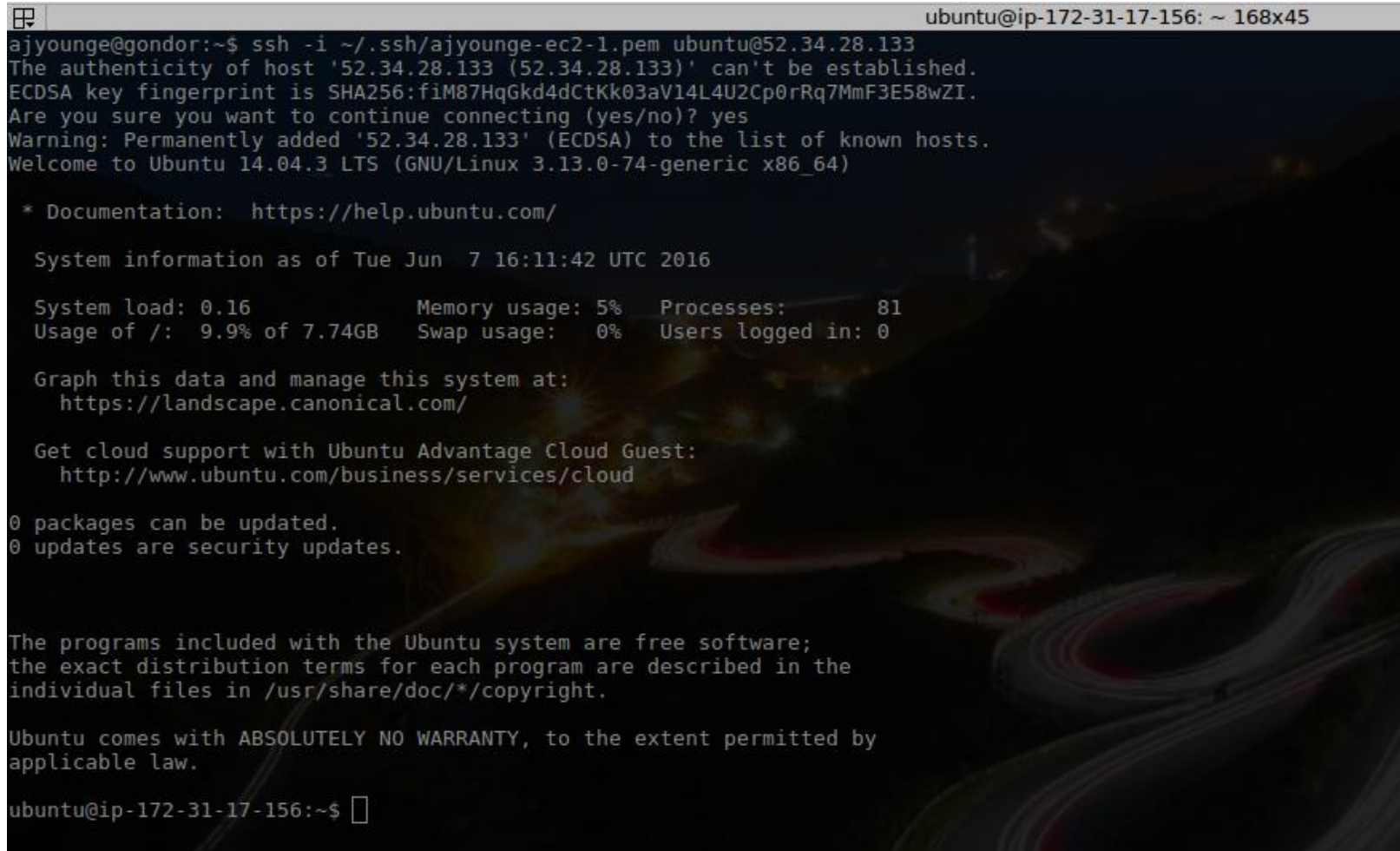
SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

39

Login via SSH to your Instance

```
# ssh -I ~/.ssh/ajyounge-ec2-1.pem ubuntu@52.34.28.133
```

A terminal window titled 'ubuntu@ip-172-31-17-156: ~ 168x45' showing the execution of an SSH command. The user 'ajyounge@gondor' runs 'ssh -i ~/.ssh/ajyounge-ec2-1.pem ubuntu@52.34.28.133'. The terminal displays a warning about the host's authenticity, a confirmation to continue, and system information for Ubuntu 14.04.3 LTS. The system info includes load, memory, processes, disk usage, swap usage, and logged-in users. It also provides links for documentation, system management, and cloud support, followed by update status and warranty information.

```
ubuntu@ip-172-31-17-156: ~ 168x45
ajyounge@gondor:~$ ssh -i ~/.ssh/ajyounge-ec2-1.pem ubuntu@52.34.28.133
The authenticity of host '52.34.28.133 (52.34.28.133)' can't be established.
ECDSA key fingerprint is SHA256:fiM87HqGkd4dCtKk03aV14L4U2Cp0rRq7MmF3E58wZI.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '52.34.28.133' (ECDSA) to the list of known hosts.
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 Tue Jun  7 16:11:42 UTC 2016

System load: 0.16           Memory usage: 5%    Processes:           81
Usage of /:  9.9% of 7.74GB  Swap usage:  0%    Users logged in:  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

0 packages can be updated.
0 updates are security updates.

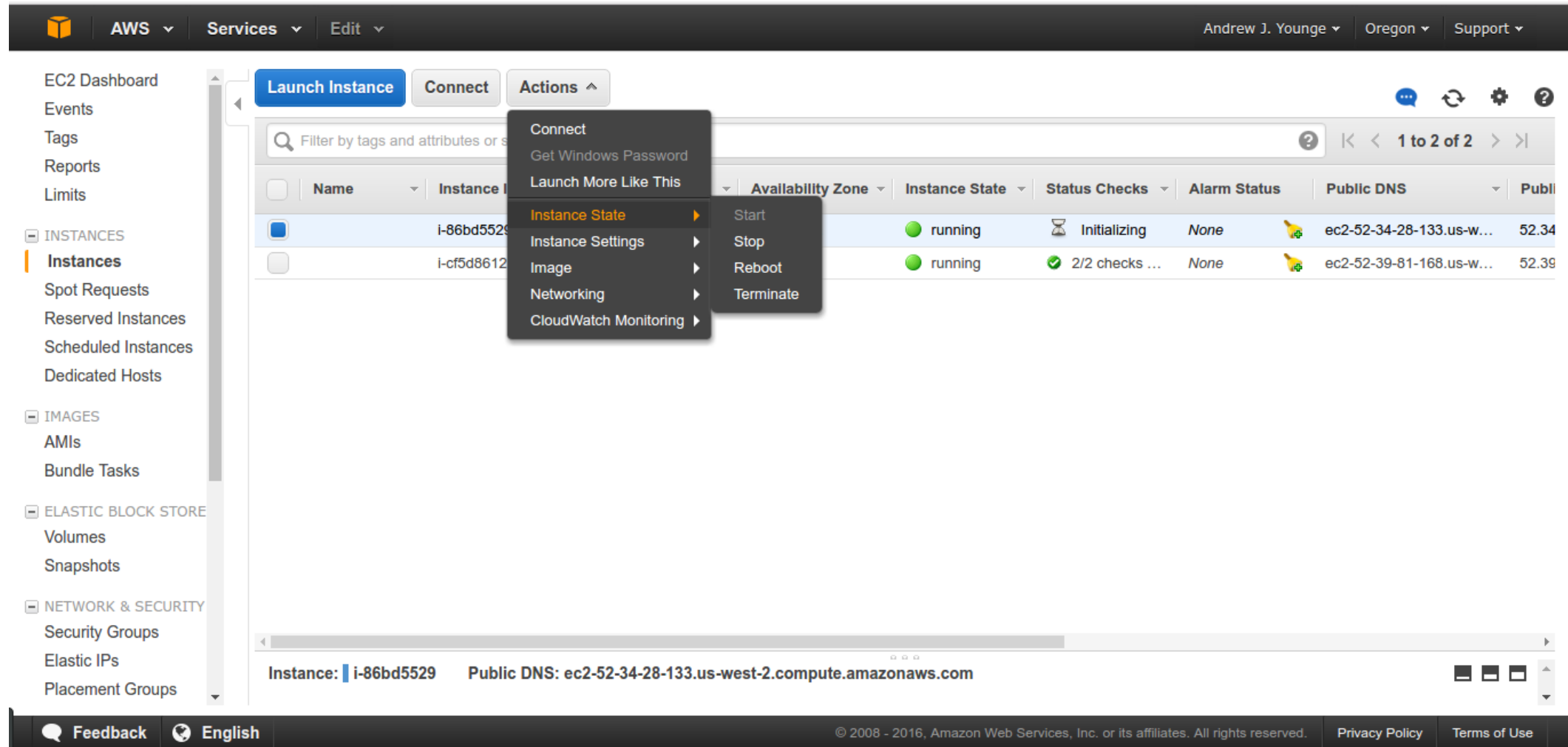
The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

ubuntu@ip-172-31-17-156:~$
```



Manage Instance State



The screenshot displays the AWS Management Console interface for managing EC2 instances. The left sidebar shows navigation options like EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, IMAGES, ELASTIC BLOCK STORE, and NETWORK & SECURITY. The main content area shows a table of instances with columns for Name, Instance ID, Availability Zone, Instance State, Status Checks, Alarm Status, Public DNS, and Public IP. Two instances are listed: i-86bd5529 (running, Initializing) and i-cf5d8612 (running, 2/2 checks passed). The 'Actions' menu is open, showing options like Connect, Get Windows Password, Launch More Like This, Instance State (highlighted), Instance Settings, Image, Networking, and CloudWatch Monitoring. The 'Instance State' sub-menu is also open, showing Start, Stop, Reboot, and Terminate.

Name	Instance ID	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
	i-86bd5529	us-west-2a	running	Initializing	None	ec2-52-34-28-133.us-west-2.compute.amazonaws.com	52.34.28.133
	i-cf5d8612	us-west-2a	running	2/2 checks passed	None	ec2-52-39-81-168.us-west-2.compute.amazonaws.com	52.39.81.168

Instance: i-86bd5529 Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com



Manage Instance Settings

The screenshot displays the AWS Management Console interface. At the top, the navigation bar includes the AWS logo, 'Services', 'Edit', and user information for 'Andrew J. Younge' in the 'Oregon' region. The left-hand navigation pane lists various services, with 'INSTANCES' expanded to show 'Instances', 'Spot Requests', 'Reserved Instances', 'Scheduled Instances', and 'Dedicated Hosts'. The main content area shows the 'Launch Instance' button, a search filter, and a table of EC2 instances. The 'Actions' dropdown menu is open for the instance 'i-86bd5529', showing options like 'Connect', 'Launch More Like This', 'Instance State', 'Instance Settings' (highlighted), 'Image', 'Networking', and 'CloudWatch Monitoring'. The 'Instance Settings' sub-menu is also open, listing actions such as 'Add/Edit Tags', 'Attach to Auto Scaling Group', 'Change Instance Type', 'Change Termination Protection', 'View/Change User Data', 'Change Shutdown Behavior', 'Get System Log', 'Get Instance Screenshot', and 'Modify Instance Placement'. The instance table shows two instances: 'i-86bd5529' in the 'us-west-2' region, 'us-east-1a' availability zone, with state 'Initializing' and public DNS 'ec2-52-34-28-133.us-west-2.compute.amazonaws.com'; and 'i-cf5d8612' in the 'us-west-2' region, 'us-east-1a' availability zone, with state 'Running' and public DNS 'ec2-52-39-81-168.us-west-2.compute.amazonaws.com'. The bottom of the console shows the instance details for 'i-86bd5529' and its public DNS.

EC2 Dashboard
Events
Tags
Reports
Limits

INSTANCES
Instances
Spot Requests
Reserved Instances
Scheduled Instances
Dedicated Hosts

IMAGES
AMIs
Bundle Tasks

ELASTIC BLOCK STORE
Volumes
Snapshots

NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups

Launch Instance Connect Actions

Filter by tags and attributes or search

Connect
Get Windows Password
Launch More Like This
Instance State
Instance Settings
Image
Networking
CloudWatch Monitoring

Add/Edit Tags
Attach to Auto Scaling Group
Change Instance Type
Change Termination Protection
View/Change User Data
Change Shutdown Behavior
Get System Log
Get Instance Screenshot
Modify Instance Placement

Name	Instance ID	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
	i-86bd5529	us-east-1a	Initializing	Initializing	None	ec2-52-34-28-133.us-west-2.compute.amazonaws.com	52.34
	i-cf5d8612	us-east-1a	Running	2/2 checks ...	None	ec2-52-39-81-168.us-west-2.compute.amazonaws.com	52.39

Instance: i-86bd5529 Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com

Feedback English © 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

Manage Instance Networking

The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, 'Services', 'Edit', and user information for Andrew J. Younge in the Oregon region. The left sidebar contains navigation links for EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, IMAGES, ELASTIC BLOCK STORE, and NETWORK & SECURITY. The main content area displays the 'Launch Instance' button, a search filter, and a table of EC2 instances. The 'Actions' menu is open for instance i-86bd5529, showing options like 'Connect', 'Launch More Like This', 'Instance State', 'Instance Settings', 'Image', 'Networking', and 'CloudWatch Monitoring'. The 'Networking' submenu is expanded, listing actions such as 'Change Security Groups', 'Attach Network Interface', 'Detach Network Interface', 'Disassociate Elastic IP Address', 'Change Source/Dest. Check', and 'Manage Private IP Addresses'. The instance table shows two instances: i-86bd5529 (us-west-2a, running, Initializing) and i-cf5d8612 (us-west-2c, running, 2/2 checks ...).

Name	Instance ID	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
	i-86bd5529	us-west-2a	running	Initializing	None	ec2-52-34-28-133.us-w...	52.34
	i-cf5d8612	us-west-2c	running	2/2 checks ...	None	ec2-52-39-81-168.us-w...	52.39

Instance: i-86bd5529 Public DNS: ec2-52-34-28-133.us-west-2.compute.amazonaws.com



Terminate your Instance

- Make sure to terminate all your instances when you are finished
- Remember: You pay by the hour
- Even small instances can rack up large bills if left running!
- NOTE: You will lose all data when you terminate an instance.
 - Backup data to EBS, S3, or personal workstation.
 - Create an image snapshot to save current file system state.

Terminate Instances

Warning

On an EBS-backed instance, the default action is for the root EBS volume to be deleted when the instance is terminated. Storage on any local drives will be lost.

Are you sure you want to terminate these instances?

- i-86bd5529 (ec2-52-34-28-133.us-west-2.compute.amazonaws.com)

Cancel

Yes, Terminate

Create Snapshot

Volume

vol-4145a7f9

Name

test-snapshot

Description

Illustrating how to create a custom snapshot

Encrypted

No

Cancel

Create

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
		i-86bd5529	t2.micro	us-west-2a	shutting-do...		None		



Hands-on 2

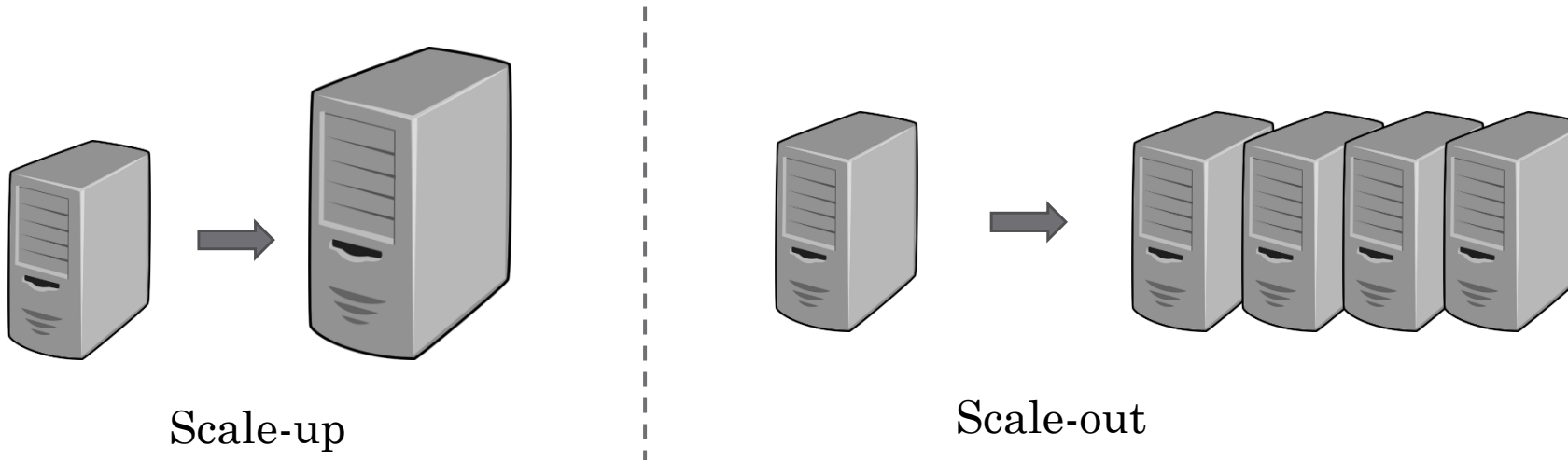
Questions?



MapReduce

- What Happened in ~2004
 - Google wanted to process web data – a whole lot of web data
 - Also, do it in a scale-out fashion over commodity hardware
 - With fault-tolerance too
 - They developed MapReduce
 - MapReduce: simplified data processing on large clusters

(<http://dl.acm.org/citation.cfm?id=1251264>)



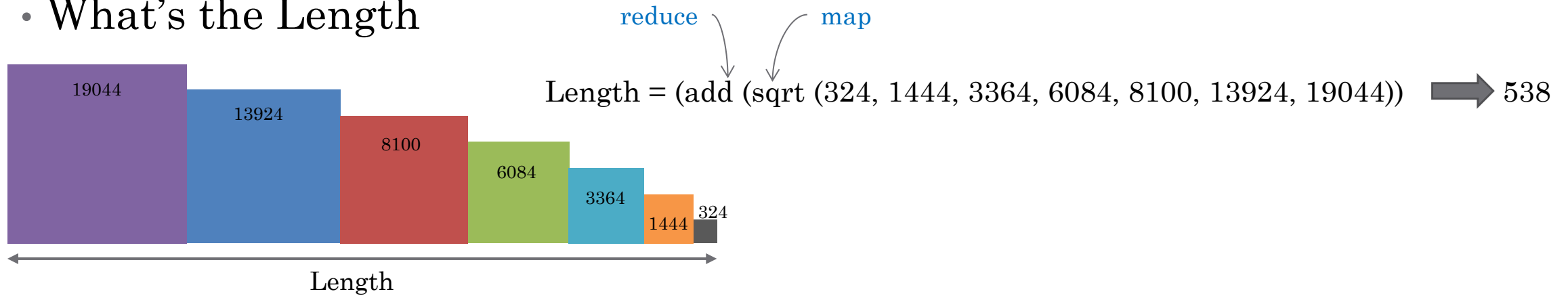
What's MapReduce

- The Concept Isn't New
 - A *list of values* mapped into another *list of values*, which gets reduced into a *single value*
 - Apply a function – *map()* – to individual data items
 - Collect results with a reduction function – *reduce()*
 - Dates back to Lambda calculus
- Google's Implementation
 - A *list of <key, value>* pairs mapped into another *list of <key, value>* pairs, which gets grouped by the key and reduced into a *list of values*
 - Distributed and horizontally scalable
 - Fault tolerant
 - Easy to program



A Few Examples

- What's the Length



- Counting Words

“*Mary* had a little *lamb*,
His fleece was white as snow,
And everywhere that *Mary* went,
The *lamb* was sure to go”



{(Mary, 2), (had, 1), (a, 1), (little, 1), (lamb, 2),
(His, 1), (fleece, 1), (was, 2), (white, 1), (as, 1),
(snow, 1), (And, 1), (everywhere, 1), (that, 1),
(went, 1), (The, 1), (sure, 1), (to, 1), (go, 1) }



Why Is It Easy?

- Think in Map and Reduce
 - Simplified abstraction – somewhat resembles Legos with just **two type of blocks**
- Hides Intricacies of Parallel Programming
 - Communication, data distribution, fault-tolerance, etc.
- Many Applications Fall into MapReduce Model and Its Extensions
 - Distributed Grep
 - Calculating Statistics
 - Page Rank
 - K-Means
 - Multidimensional Scaling
 - See
 - <http://web.cs.wpi.edu/~cs4513/d08/OtherStuff/MapReduce-TeamC.ppt>
 - Many other applications, if you Google ☺



Apache Hadoop (It's Free !!)

- The Open Source MapReduce Implementation
- Scalable
 - Almost linear scaling with cores and disks
 - Can handle thousands of nodes across multiple racks
 - Can handle large loads without crashing!
- Reliable
 - All the data blocks are replicated
 - Data recoverability
 - Nodes can join or leave cluster any time
- Fault Tolerance
 - Re-execution of failed tasks
 - Retry data transmissions
 - Can tolerate Hardware failures
- Simple
 - Simple storage and programming model

Hadoop MapReduce v2 Cookbook Second Edition

https://www.amazon.com/Hadoop-MapReduce-v2-Cookbook-Second-ebook/dp/B00U1D9WT6?ie=UTF8&ref=asap_bc



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

Apache Hadoop

- Distributed Storage (HDFS)
 - Not LUSTRE or a SAN..
 - Can't do random reads/writes
 - But cheap, reliable and scalable
 - Parallel storage
 - Very large aggregate bandwidth
- Processing
 - Not MPI
 - Can't do inter process communication or collective operations
 - But highly scalable, easy to program and runs on commodity hardware
 - Fault tolerant, dynamic scheduling
- Querying and Table storage
 - Not Netezza or Teradata
 - Do not support full SQL, full indexing and has high latency
 - But highly scalable, cheap and fast for very large data sets



Why Hadoop?

- Not the best in any of them (may be in cheap storage), but good at all of those. Taken altogether makes it very attractive.
 - Not the fastest, but scalable
 - Easy to code
 - Cheap to scale
 - Runs on commodity hardware
 - Can handle very very large data and computations
 - Battle tested in thousands of clusters
- Large open source echo system
 - Many projects add functionalities on top of HDFS and Hadoop
 - Large community of developers and users



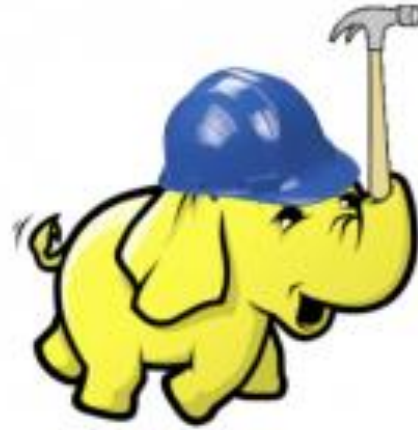
Hadoop Usage

- Yahoo!, Facebook, Netflix, Amazon, Twitter, LinkedIn, Link Analytics
- Support by Cloudera, Hortonworks, Intel, IBM, MapR, etc.
- Processing petabytes of data daily
- Yahoo Hadoop cluster is 40,000 nodes
- Facebook is storing more than 100PB in their Hadoop cluster
- Hosted Hadoop as a service by Amazon EMR, Microsoft Azure, Google..



Hadoop is Not!

- Hadoop is a very big Hammer!
 - Not for small data / jobs
 - Not to store ton of small files
 - Real-time or interactive results
 - For hard to parallelize problems



Apache Big Data Stack

- More Than Hadoop
- Over 350 Open Source Software Packages
 - As of January 2016
- Popular Projects
 - Apache Hadoop
 - Apache Storm
 - Apache Spark
 - Apache Flink



Cross-Cutting Functions	17) Workflow-Orchestration: ODE, ActiveBPEL, Airavata, Pegasus, Kepler, Swift, Taverna, Triana, Trident, BioKepler, Galaxy, IPython, Dryad, Naiad, Oozie, Tez, Google FlumeJava, Crunch, Cascading, Scalding, e-Science Central, Azure Data Factory, Google Cloud Dataflow, NiFi (NSA), Jitterbit, Talend, Pentaho, Apatar, Docker Compose, KeystoneML
1) Message and Data Protocols: Avro, Thrift, Protobuf	16) Application and Analytics: Mahout , MLlib , MLbase, DataFu, R, pbdR, Bioconductor, ImageJ, OpenCV, Scalapack, PetSc, PLASMA MAGMA, Azure Machine Learning, Google Prediction API & Translation API, mipy, scikit-learn, PyBrain, CompLearn, DAAL(Intel), Caffe, Torch, Theano, DL4j, H2O, IBM Watson, Oracle PGX, GraphLab, GraphX, IBM System G, GraphBuilder(Intel), TinkerPop, Parasol, Dream:Lab, Google Fusion Tables, CINET, NWB, Elasticsearch, Kibana, Logstash, Graylog, Splunk, Tableau, D3.js, three.js, Potree, DC.js, TensorFlow, CNTK
2) Distributed Coordination: Google Chubby, Zookeeper, Giraffe, JGroups	15B) Application Hosting Frameworks: Google App Engine, AppScale, Red Hat OpenShift, Heroku, Aerobatic, AWS Elastic Beanstalk, Azure, Cloud Foundry, Pivotal, IBM BlueMix, Ninefold, Jelastic, Stackato, appfog, CloudBees, Engine Yard, CloudControl, dotCloud, Dokku, OSGi, HUBzero, OODT, Agave, Atmosphere 15A) High level Programming: Kite, Hive, HCatalog, Tajo, Shark, Phoenix, Impala, MRQL, SAP HANA, HadoopDB, PolyBase, Pivotal HD/Hawq, Presto, Google Dremel, Google BigQuery, Amazon Redshift, Drill, Kyoto Cabinet, Pig, Sawzall, Google Cloud DataFlow, Summingbird
3) Security & Privacy: InCommon, Eduroam, OpenStack, Keystone, LDAP, Sentry, Sqrrl, OpenID, SAML OAuth	14B) Streams: Storm, Samza, Granules, Neptune, Google MillWheel, Amazon Kinesis, LinkedIn, Twitter Heron, Databus, Facebook Puma/Ptail/Scribe/ODS, Azure Stream Analytics, Floe, Spark Streaming, Flink Streaming, DataTurbine 14A) Basic Programming model and runtime, SPMD, MapReduce: Hadoop, Spark, Twister, MR-MPI, Stratosphere (Apache Flink), Reef, Disco, Hama, Giraph, Pregel, Pegasus, Ligra, GraphChi, Galois, Medusa-GPU, MapGraph, Totem 13) Inter process communication Collectives, point-to-point, publish-subscribe: MPI, HPX-5, Argo BEAST HPX-5 BEAST PULSAR, Harp, Netty, ZeroMQ, ActiveMQ, RabbitMQ, NaradaBrokering, QPid, Kafka, Kestrel, JMS, AMQP, Stomp, MQTT, Marionette Collective, Public Cloud: Amazon SNS, Lambda, Google Pub Sub, Azure Queues, Event Hubs 12) In-memory databases/caches: Gora (general object from NoSQL), Memcached, Redis, LMDB (key value), Hazelcast, Ehcache, Infinispan, VoltDB, H-Store 12) Object-relational mapping: Hibernate, OpenJPA, EclipseLink, DataNucleus, ODBC/JDBC 12) Extraction Tools: UIMA, Tika
4) Monitoring: Ambari, Ganglia, Nagios, Inca	11C) SQL(NewSQL): Oracle, DB2, SQL Server, SQLite, MySQL, PostgreSQL, CUBRID, Galera Cluster, SciDB, Rasdaman, Apache Derby, Pivotal Greenplum, Google Cloud SQL, Azure SQL, Amazon RDS, Google F1, IBM dashDB, N1QL, BlinkDB, Spark SQL
21 layers Over 350 Software Packages January 29 2016	11B) NoSQL: Lucene, Solr, Solandra, Voldemort, Riak, ZHT, Berkeley DB, Kyoto/Tokyo Cabinet, Tycoon, Tyrant, MongoDB, Espresso, CouchDB, Couchbase, IBM Cloudant, Pivotal Gemfire, HBase, Google Bigtable, LevelDB, Megastore and Spanner, Accumulo, Cassandra, RYA, Sqrrl, Neo4J, graphdb, Yarcdata, AllegroGraph, Public Cloud: Azure Table, Amazon Dynamo, Google DataStore 11A) File management: iRODS, NetCDF, CDF, HDF, OPeNDAP, FITS, RCFile, ORC, Parquet 10) Data Transport: BitTorrent, HTTP, FTP, SSH, Globus Online (GridFTP), Flume, Sqoop, Pivotal GPLOAD/GPFDIST 9) Cluster Resource Management: Mesos, Yarn, Helix, Llama, Google Omega, Facebook Corona, Celery, HTCondor, SGE, OpenPBS, Moab, Slurm, Torque, Globus Tools, Pilot Jobs 8) File systems: HDFS, Swift, Haystack, f4, Cinder, Ceph, FUSE, Gluster, Lustre, GPFS, GFFS Public Cloud: Amazon S3, Azure Blob, Google Cloud Storage 7) Interoperability: Libvirt, Libcloud, JClouds, TOSCA, OCCl, CDMI, Whirr, Saga, Genesis 6) DevOps: Docker (Machine, Swarm), Puppet, Chef, Ansible, SaltStack, Boto, Cobbler, Xcat, Razor, CloudMesh, Juju, Foreman, OpenStack Heat, Sahara, Rocks, Cisco Intelligent Automation for Cloud, Ubuntu MaaS, Facebook Tupperware, AWS OpsWorks, OpenStack IroniC, Google Kubernetes, Buildstep, Gitreceive, OpenTOSCA, Winery, CloudML, Blueprints, Terraform, DevOpSlang, Any2Api 5) IaaS Management from HPC to hypervisors: Xen, KVM, QEMU, Hyper-V, VirtualBox, OpenVZ, LXC, Linux-Vserver, OpenStack, OpenNebula, Eucalyptus, Nimbus, CloudStack, CoreOS, rkt, VMware ESXi, vSphere and vCloud, Amazon, Azure, Google and other public Clouds Networking: Google Cloud DNS, Amazon Route 53



Tools of the Trade

- Programming Languages
 - Java is the dominant one in Big Data space
 - Python, C/C++ to follow
- Integrated Development Environments
 - Eclipse <https://eclipse.org/downloads/>
 - IntelliJIDEA <https://www.jetbrains.com/idea/> (personal preference)
 - Good news! The commercial version is free for students and educators
 - Both these are pretty powerful – comparing one vs the other is like Mercedes vs BMW
- Other Tools
 - Version controlling systems – Git/GitHub is currently preferred by many, so is SVN
 - Build tools – Apache Maven, Apache ANT, and
 - Testing (JUnit), Continuous Integration (CI) – Travis



When I am Stuck

- Google
 - This has become an art in its own right
- Stack Overflow
 - Works best if you know what you are trying to solve like a specific exception
- Quora
 - Trending place to ask general questions – “I am 20 I need to be a millionaire by 25. How to?”
- Learning
 - Linux – Software Carpentry <http://software-carpentry.org/> is good
 - Java – Tutorialspoint <http://www.tutorialspoint.com/java/>
 - Online courses – so many available – look in Coursera, Lynda, etc. YouTube too!



Hands-on 3

Getting Started with Apache Hadoop

Refer to

<http://admicloud.github.io/www/SetUpHadoop.html>



Programming with MapReduce

- Word Count
 - Count the occurrence of words in a set of text files
 - The de-factor “Hello, World” application of cloud computing
- K-Means
 - Given N points, group them into K clusters
 - A commonly used machine learning algorithm
- Page Rank
 - Given an adjacency matrix representing Web pages and their target pages, compute a rank for each page
 - The rank indicates the probability of someone visiting a given page, i.e. higher the rank the higher the chances it being visited by a user
 - The foundation of Google’s search algorithm



Word Count

- Input

- “Mary had a little lamb,
His fleece was white as snow,
And everywhere that Mary went,
The lamb was sure to go”

- Output

- {(Mary, 2), (had, 1), (a, 1), (little, 1), (lamb, 2), (His, 1), (fleece, 1), (was, 2), (white, 1), (as, 1), (snow, 1), (And, 1), (everywhere, 1), (that, 1), (went, 1), (The, 1), (sure, 1), (to, 1), (go, 1) }



Serial Implementation

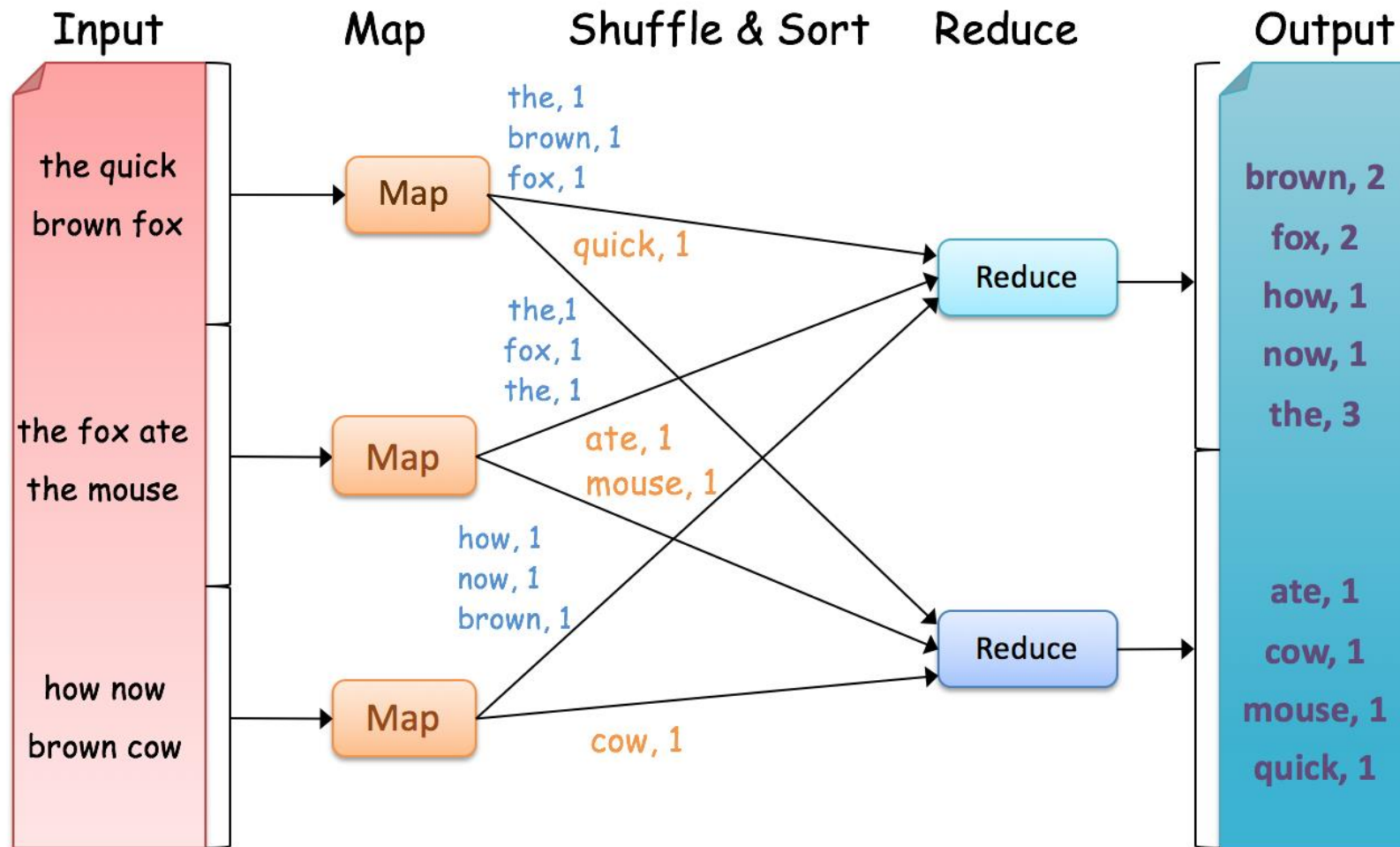
- Create a hash table (HT)
- While more lines to read
 - Read line
 - Split into words
 - For each word
 - If HT has word increment its count
 - Else add word to HT with count=1
- Output HT

```
BufferedReader br = new BufferedReader(new FileReader(wordFile));
Hashtable<String, Integer> wordToCountTable = new Hashtable<>();
Pattern pat = Pattern.compile(" ");
String line;
String [] splits;
while ((line = br.readLine()) != null) {
    splits = pat.split(line);
    for (String s:splits) {
        if (wordToCountTable.containsKey(s)) {
            wordToCountTable.put(s, wordToCountTable.get(s)+1);
            continue;
        }
        wordToCountTable.put(s, 1);
    }
}

Enumeration<String> words = wordToCountTable.keys();
String key;
while(words.hasMoreElements()) {
    key = words.nextElement();
    System.out.println(key + " " + wordToCountTable.get(key));
}
```



Hadoop (MapReduce) Implementation



Hands-on 4

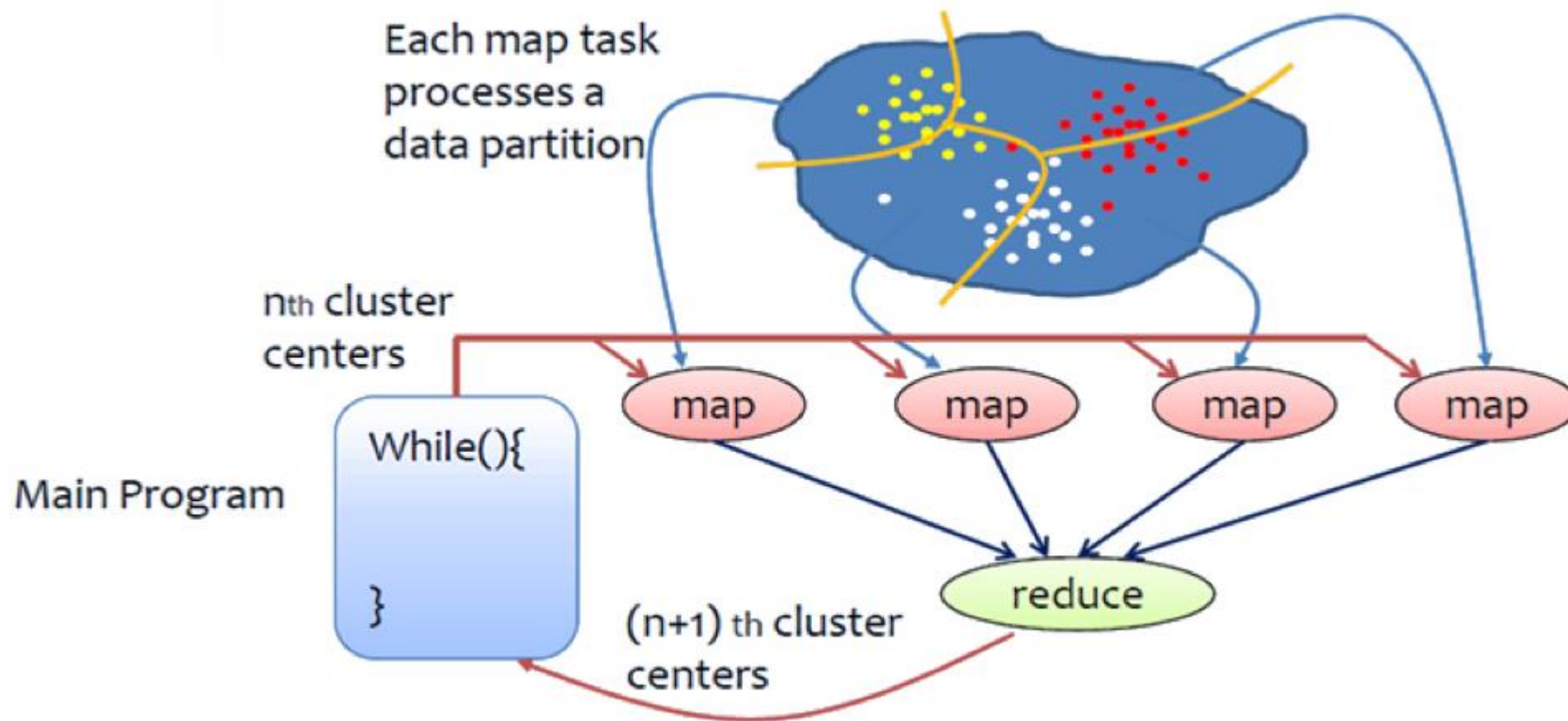
Word Count with Apache Hadoop

Refer to

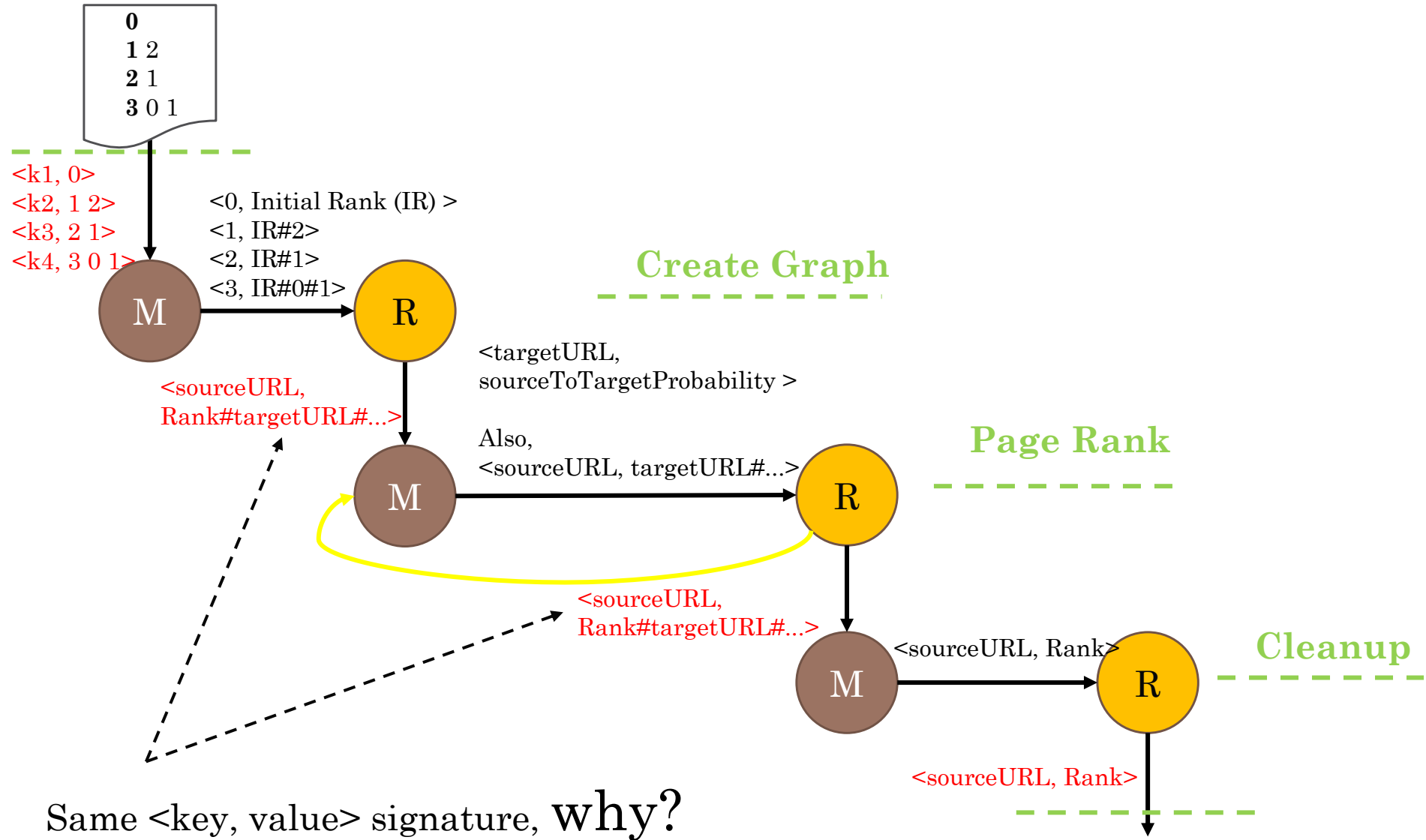
<http://admicloud.github.io/www/wordcount.html>



K-Means



Page Rank



Same $\langle \text{key, value} \rangle$ signature, why?

Output total rank sum



Hands-on 5

K-Means with Apache Hadoop

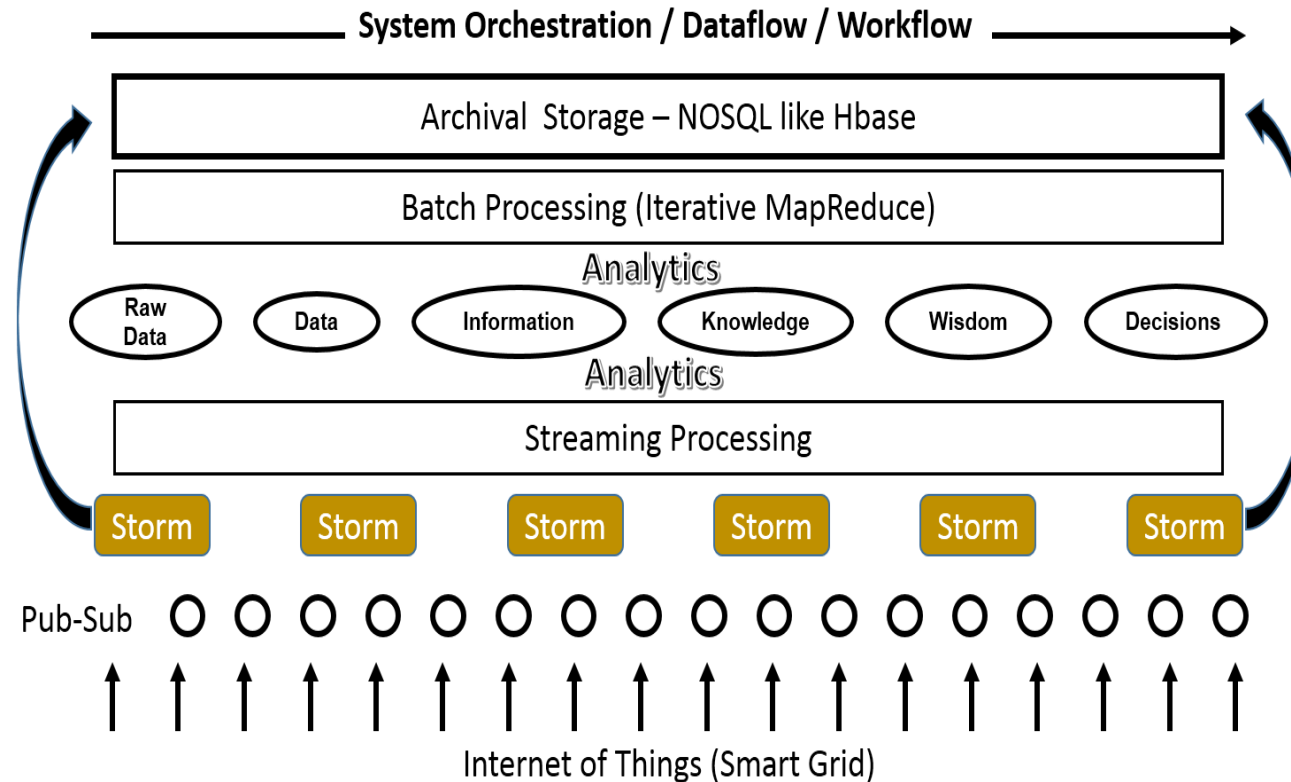
Refer to

<http://admicloud.github.io/www/kmeans.html>

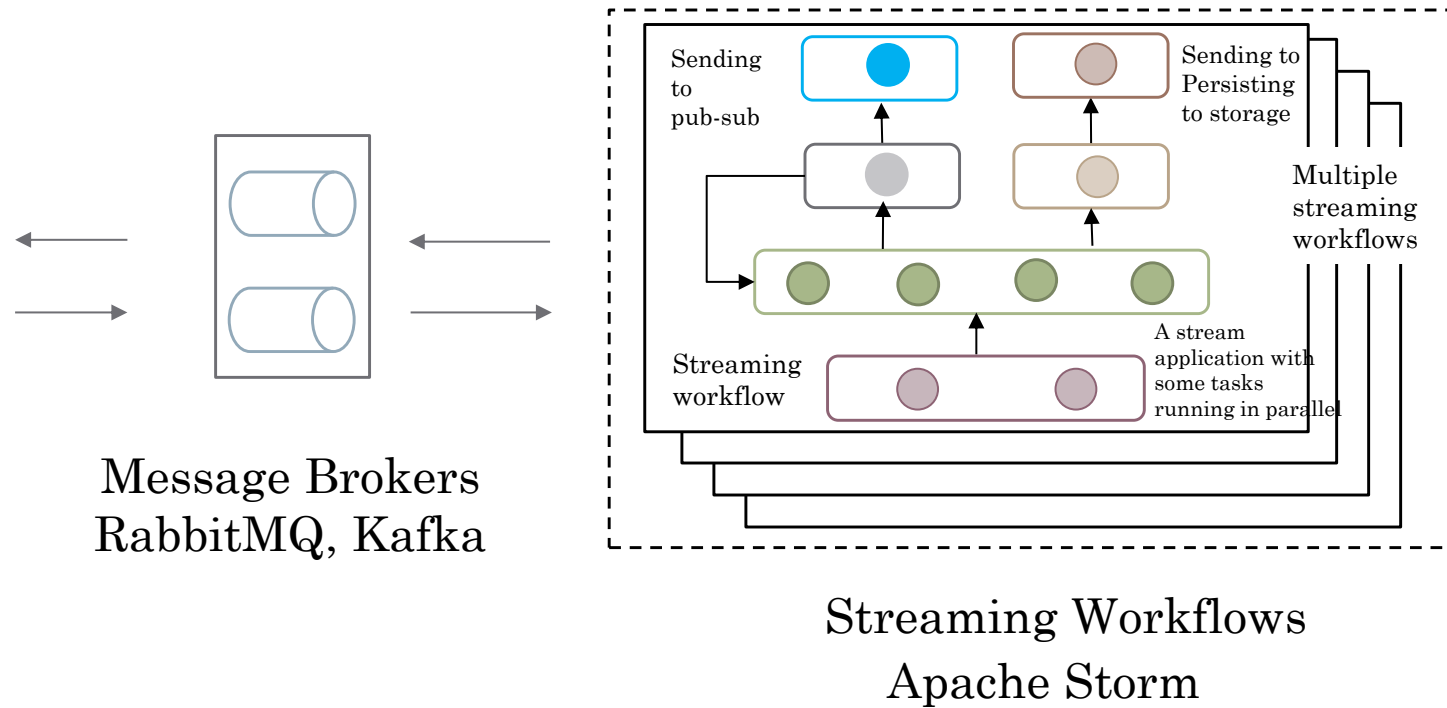


Stream Processing

- Data, Information, Knowledge, Wisdom

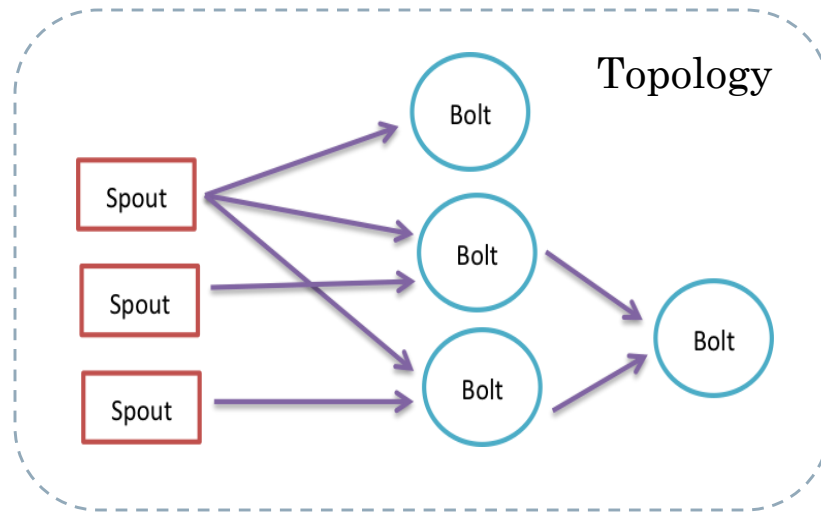


Data pipeline

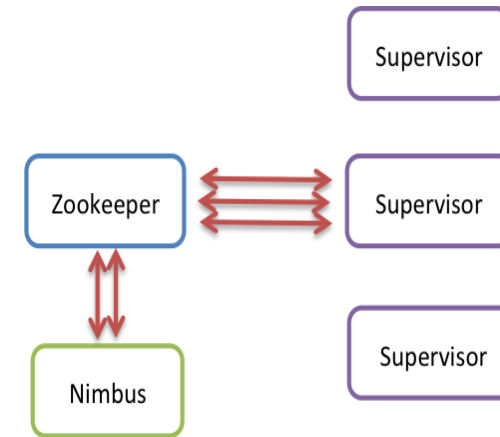


Apache Storm

- Storm is the Hadoop for distributed stream processing?
- Storm is Stream Partitioning + Fault Tolerance + Parallel Execution



Programming Model



Architecture

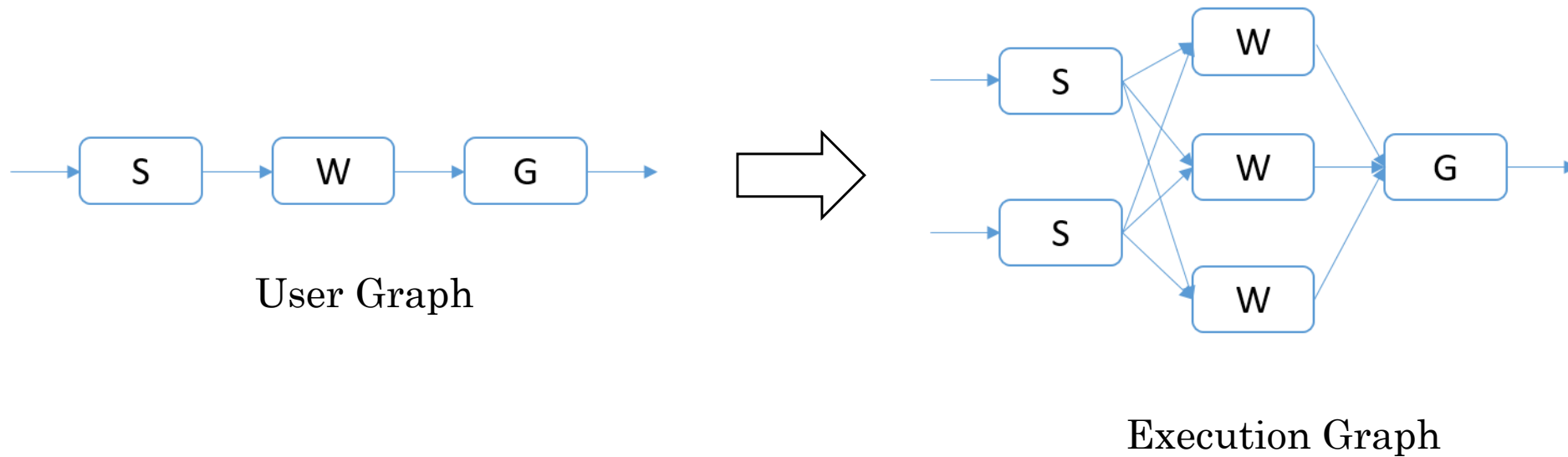
Java, Ruby, Python, Javascript, Perl, and PHP



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

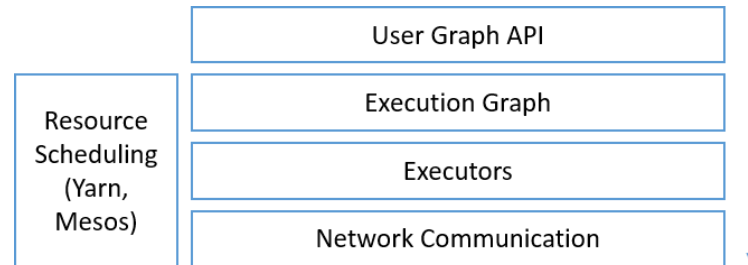
Storm Application



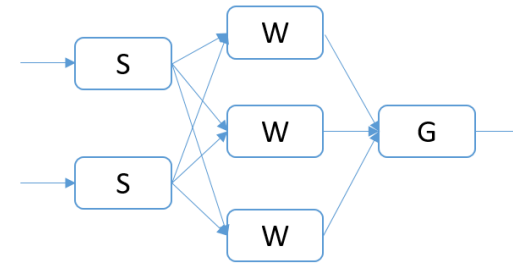
User graph is converted to an execution graph



DSPF Architecture



User graph



Execution graph

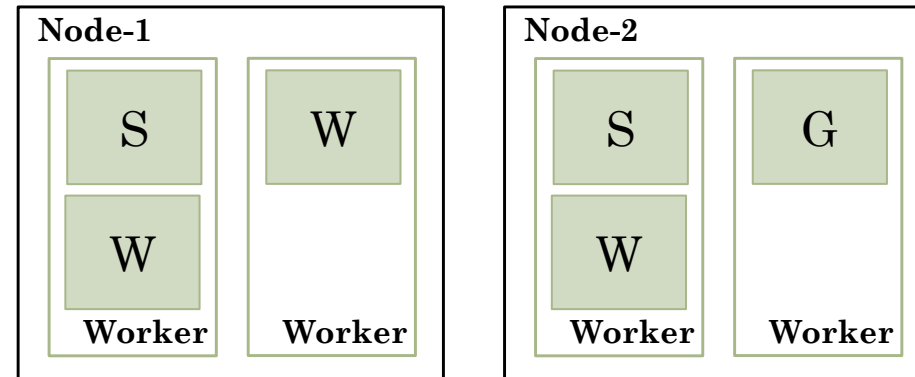


Apache Storm

- Data Mobility
 - Pull based, No blocking operations, ZeroMQ and Netty Based communication
- Fault Tolerance
 - Rollback Recovery with Upstream backup
 - The messages are saved in out queue of Spout until acknowledged
- Stream Partition
 - User defined, based on the grouping
- Storm Query Model
 - Trident, A Java library providing high level abstraction



Execution Graph Distribution in the Cluster



Two node cluster each running two workers. The tasks of the Topology is assigned to the workers



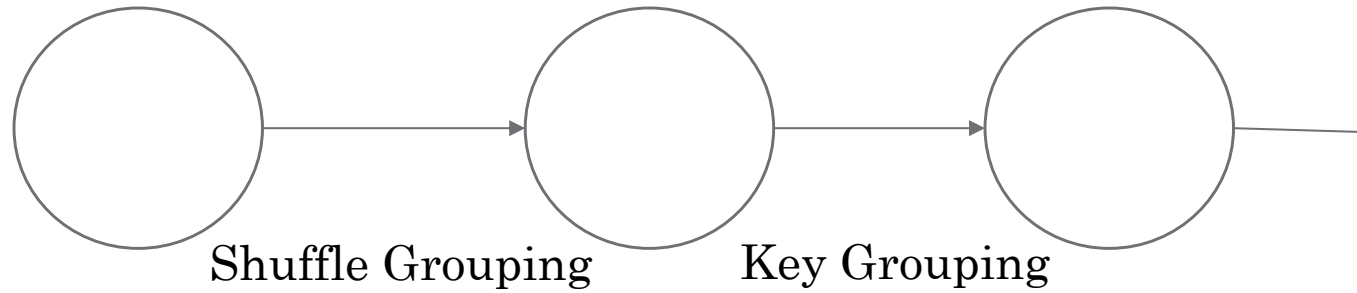
Word Count

User Topology

Sentence Generation

Split Words

Count Words



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

75

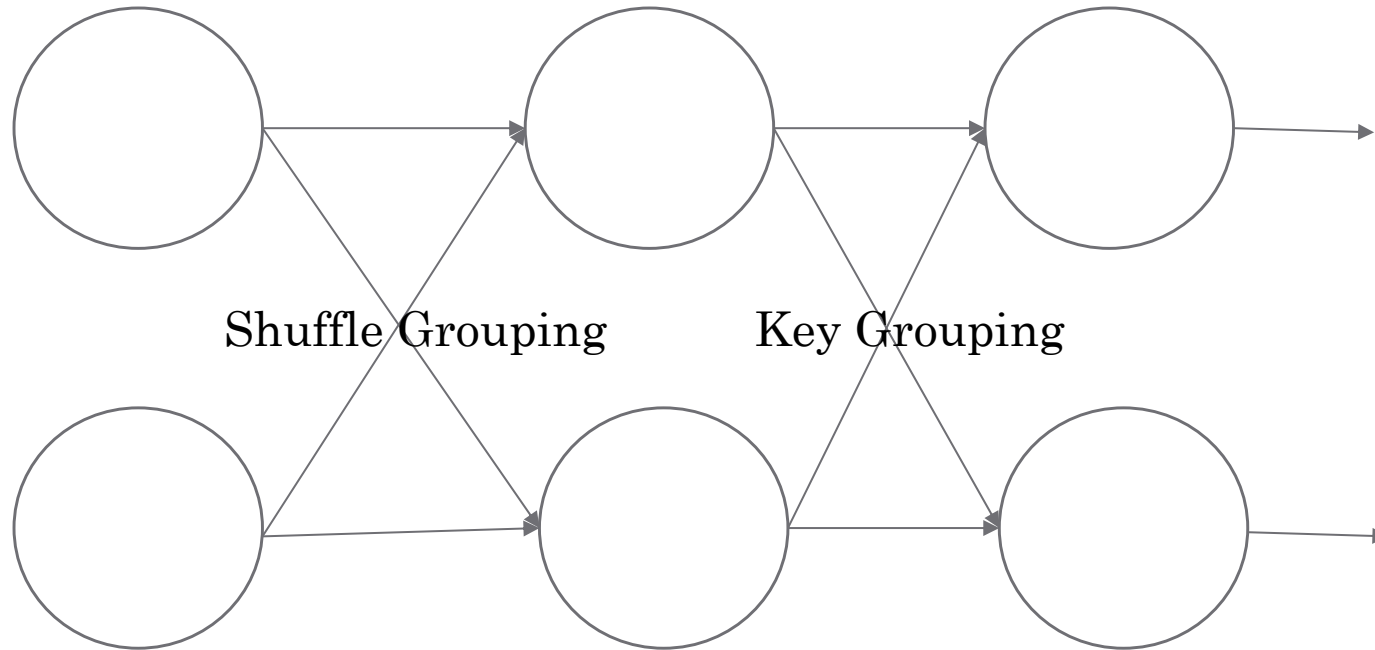
Word Count

Execution Graph

Sentence Generation

Split Words

Count Words



INDIANA UNIVERSITY BLOOMINGTON

SCHOOL OF INFORMATICS AND COMPUTING

6/10/2016

76

Hands-on 6

Streaming Word Count with Apache Storm

Refer to

<http://admicloud.github.io/www/storm.html>



Acknowledgement

- This presentation would not have been possible if not for the support of many others at IU.
- Thank you,



Judy Qiu



Andrew Younge



Ethan Li



Pulasthi Wickramasinghe



Supun Kamburugamuve



Zou, Yiming



Thomas Wiggins



Assignment: Distributed Grep with Hadoop

- Just Like Word Count
 - Except now match a given pattern
 - Output 1 only if the current word matches the pattern

