

Cloud Computing and Big Data Analytics for Teaching & Research

Qusay H. Mahmoud, Ph.D., P.Eng.

Professor & Chair

Department of Electrical, Computer and Software Engineering

University of Ontario Institute of Technology

Oshawa, ON, Canada

Agenda

2

□ Cloud Computing

- ▣ What is it? and Why?
- ▣ Service models and cloud providers
- ▣ Hands-on example: using the cloud
- ▣ Security and privacy issues
- ▣ Teaching and Research

□ Big Data Analytics

- ▣ What is it Big Data? Where does it come from?
- ▣ Technologies for processing large data sets (analytics)
- ▣ Hands-on example: processing big data
- ▣ Teaching and Research

Learning Outcomes

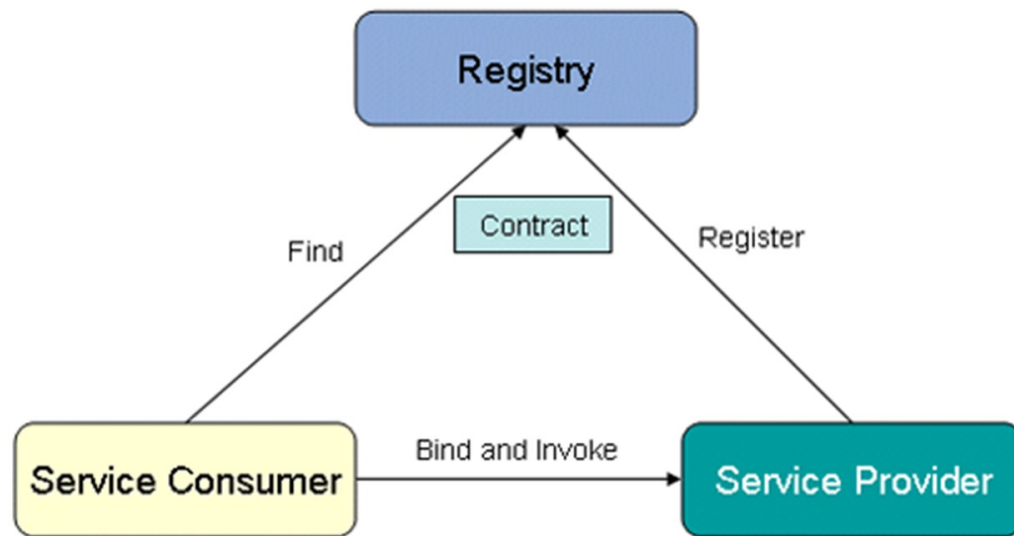
3

- At the end of this tutorial, participants will:
 - ▣ Demonstrate an understanding of cloud computing principles, service models, and provider options
 - ▣ Learn about the interesting apps of cloud and big data
 - ▣ Start using the cloud for various tasks
 - ▣ Utilize cloud computing for teaching and research
 - ▣ Learn about the research opportunities and security challenges
 - ▣ Understand the applications of cloud computing for big data, and some of the technologies for big data analytics

The Network is the Computing

4

- Service-Oriented Architecture/Computing
- SOA is an architectural style for building apps
- SOC is a computing paradigm that utilizes services as fundamental elements for developing apps



Software-as-a-Service

5

- Enterprise software is being transformed from an installed product to a hosted service



- Customer pays on a subscription or pay per use basis to access functionality using a web browser or other clients
- Corel tried this in 1997 with WordPerfect as a Java Applet

Software-as-a-Service

6



□ Benefits

- ▣ Reduced acquisition and maintenance costs (customer)
- ▣ Scalability and QoS is responsibility of service provider (customer)
- ▣ Easier support and maintenance (service provider)

□ Downside

- ▣ Security: confidential info visible to others (customer)
- ▣ Availability and reliability are harder to guarantee (service provider)



Introduction to Cloud Computing

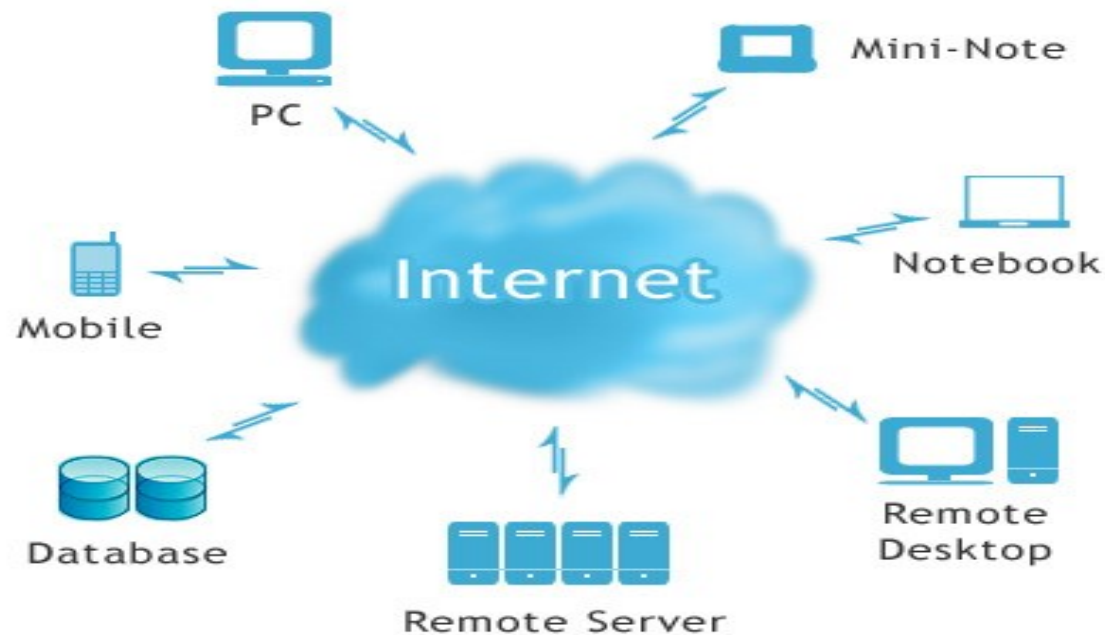
7

- NIST Definition of Cloud Computing
 - NIST: National Institute of Standards and Technology
 - Cloud computing is a model for enabling 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
 - This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models. ...

Why is it called the 'cloud'?

8

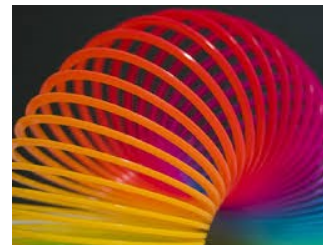
- The term derives from the fact that most technology diagrams depict the Internet by using a drawing of a cloud



Five Characteristics

9

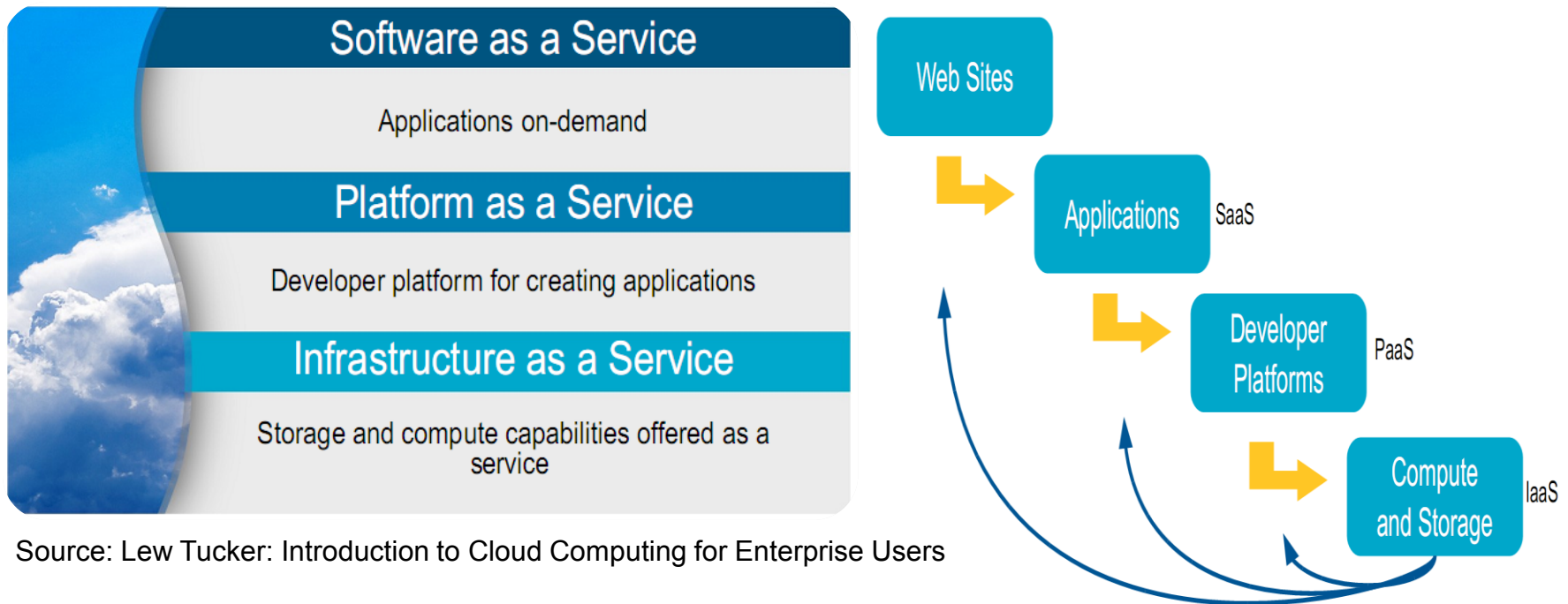
- ❑ On-demand self-service
 - ❑ It's there when you need it
- ❑ Broad network access
 - ❑ Connectivity options
- ❑ Resource pooling
 - ❑ Sharing resources (undisclosed location)
- ❑ Rapid elasticity
 - ❑ You get what you need
- ❑ Measured service
 - ❑ You pay for what you use



Three Service Models

10

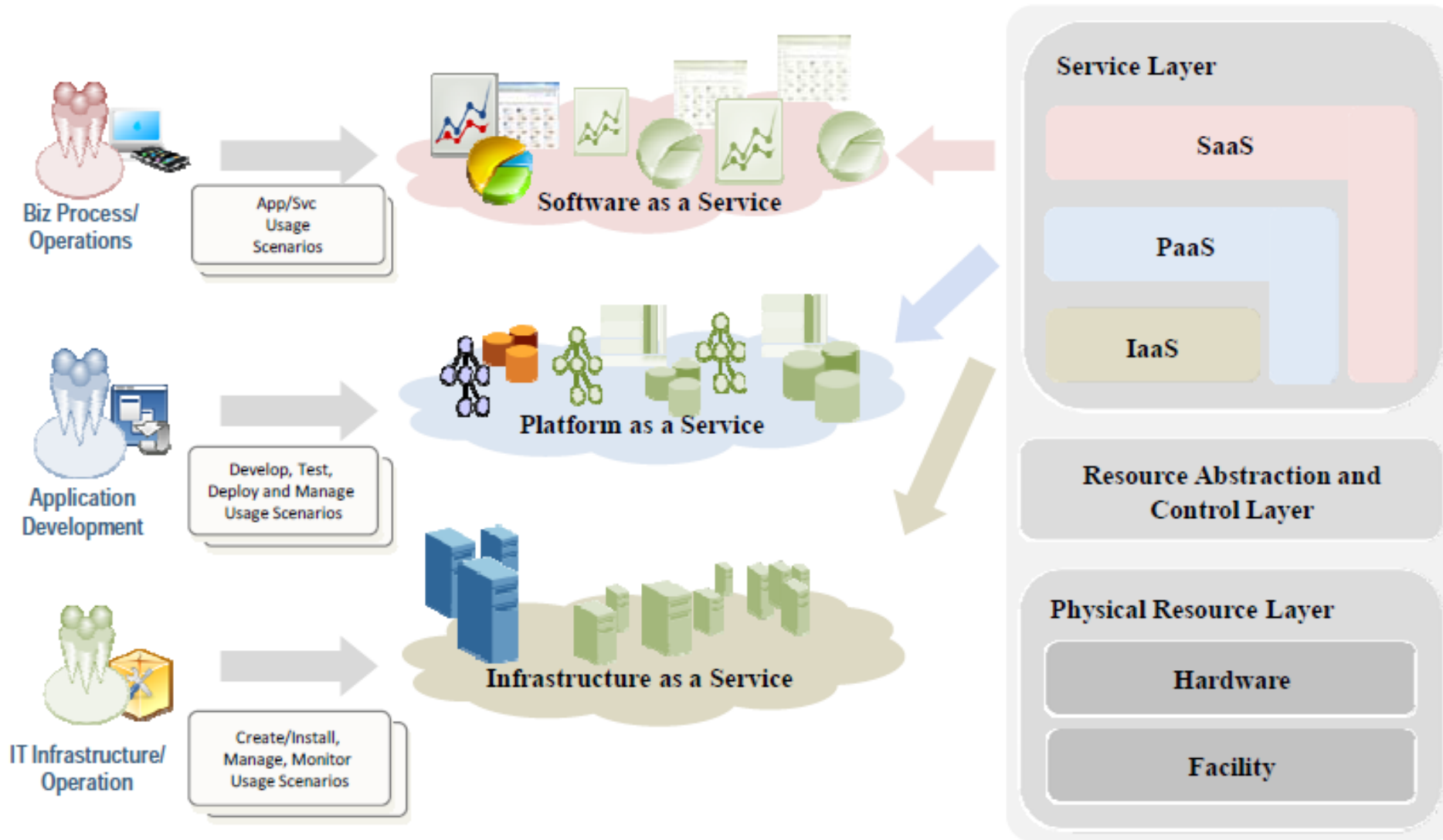
- ❑ Software as a service (SaaS)
- ❑ Platform as a service (PaaS)
- ❑ Infrastructure as a service (IaaS)



Source: Lew Tucker: Introduction to Cloud Computing for Enterprise Users

Service Orchestration

11



Cloud Computing Landscape

12

- Some of the cloud providers (or players in this space)

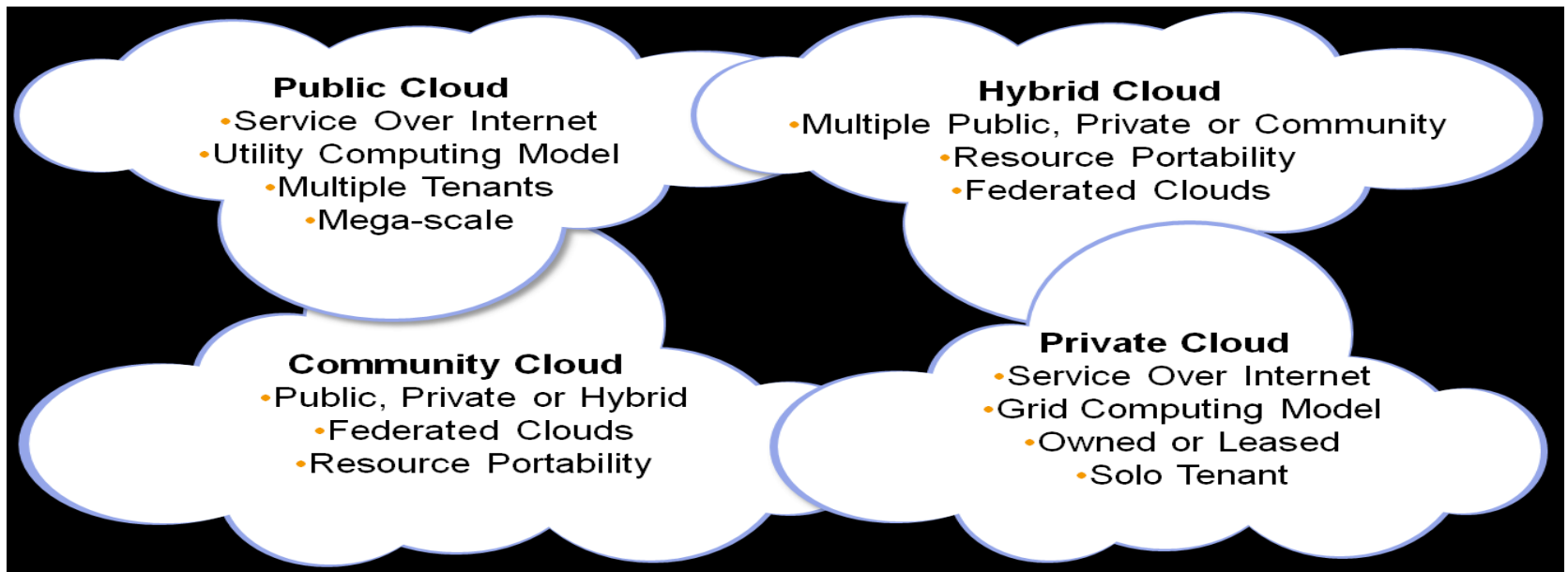


Source: Google images

Four Deployment Models

13

- ❑ Private cloud
- ❑ Community cloud
- ❑ Public cloud
- ❑ Hybrid cloud



Source: Google images

Utility Computing Business Model

14

- Cloud computing...the 5th utility?



Software as a Service (SaaS)

15

- A software delivery model in which software and its associated data are hosted centrally (typically in the cloud) and are accessed by users via a thin client, normally a web browser over the Internet

- Many services available

Linked in

EVERNOTE

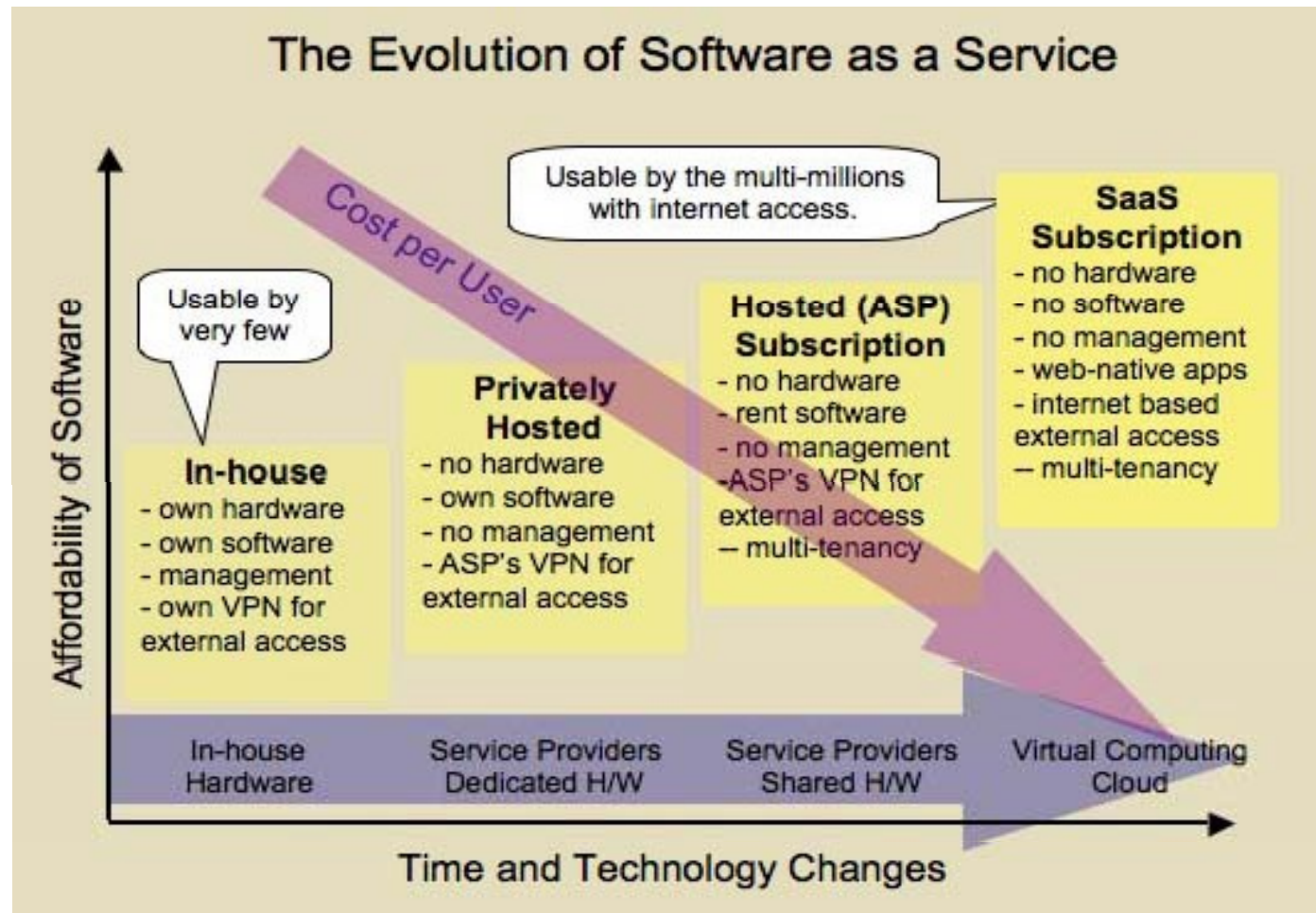


- Requirements: customizability and user data hosting

SaaS

16

□ SaaS evolution



Cloud Economics

17

- Migrating all UOIT students to Gmail

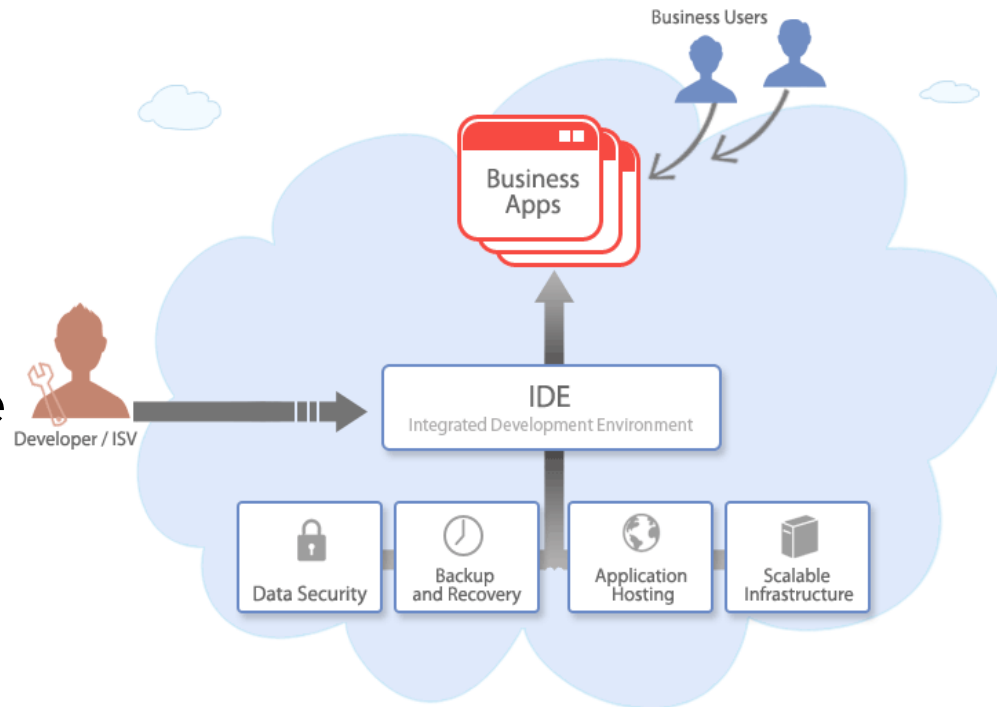


- uoit.ca email (faculty & staff – not Gmail)
- uoit.net email (student & alumni accounts, Gmail)

Platform as a Service (PaaS)

18

- A supporting layer allowing users and developers to focus on their tasks
- Provide development, testing, deployment, hosting and maintenance solutions for cloud applications



SaaS Provider: Heroku

19

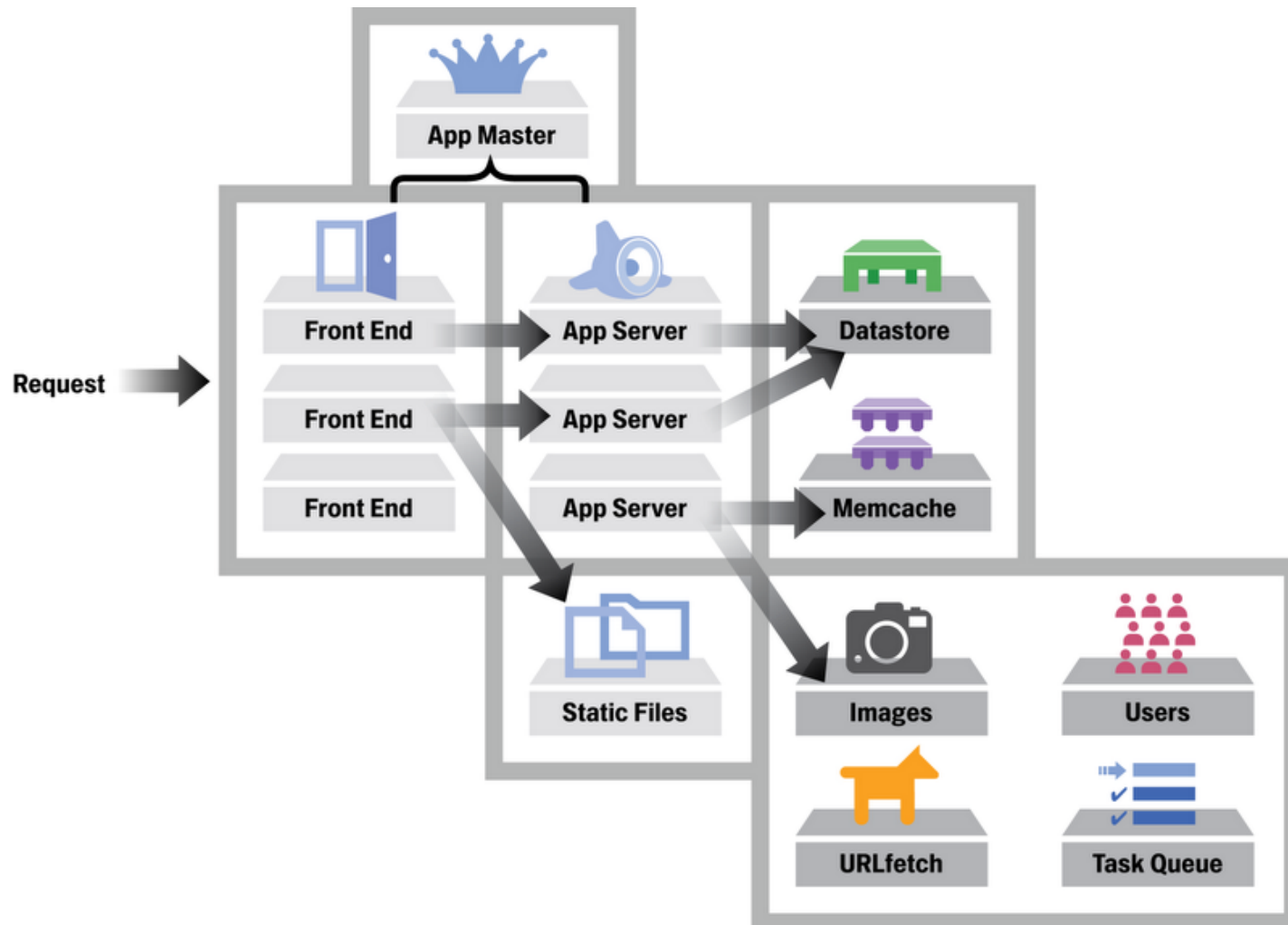
- Heroku
 - ▣ Application platform



- ▣ 12-factor app (12factor.net)
- ▣ Online demo
 - Oppia.herokuapp.com

PaaS Provider: Google App Engine

20



Google App Engine

21

- App Master
 - ▣ Schedules applications
 - ▣ Manage the replication of the applications
- Front Ends
 - ▣ Route dynamic requests to
 - Static files: if accessing static web pages
 - App Servers: if accessing dynamic contents
 - ▣ Load balancing
 - Select the nearest and lightest-loaded server (for both static and dynamic contents)

Google App Engine

22

□ App server

- ▣ Customer can provide App server logic and deploy it in Google App Engine (Everything else is automated)
- ▣ Each App server has an isolated execution environment
- ▣ Can invoke APIs to do some tasks easily



- ❖ Mail: APIs to gmail
- ❖ Users: APIs to google user account info
- ❖ Image: APIs to manipulate images, resize, crop, ...
- ❖ URLfetch: fetch other URLs
- ❖ Task Queue: support multiple threads in App, allow it to perform background tasks while handling user request
- ❖ XMPP: APIs to google talk

Hands-on Example

23

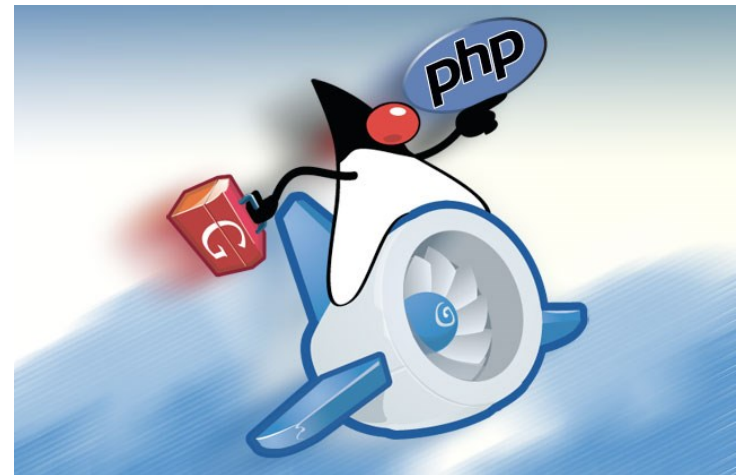
□ Developing a simple SaaS app:

▣ Google App Engine

■ Java, Python, Go

▣ Heroku (if time allows)

■ Supports many languages
and frameworks



Characteristics of Cloud Computing

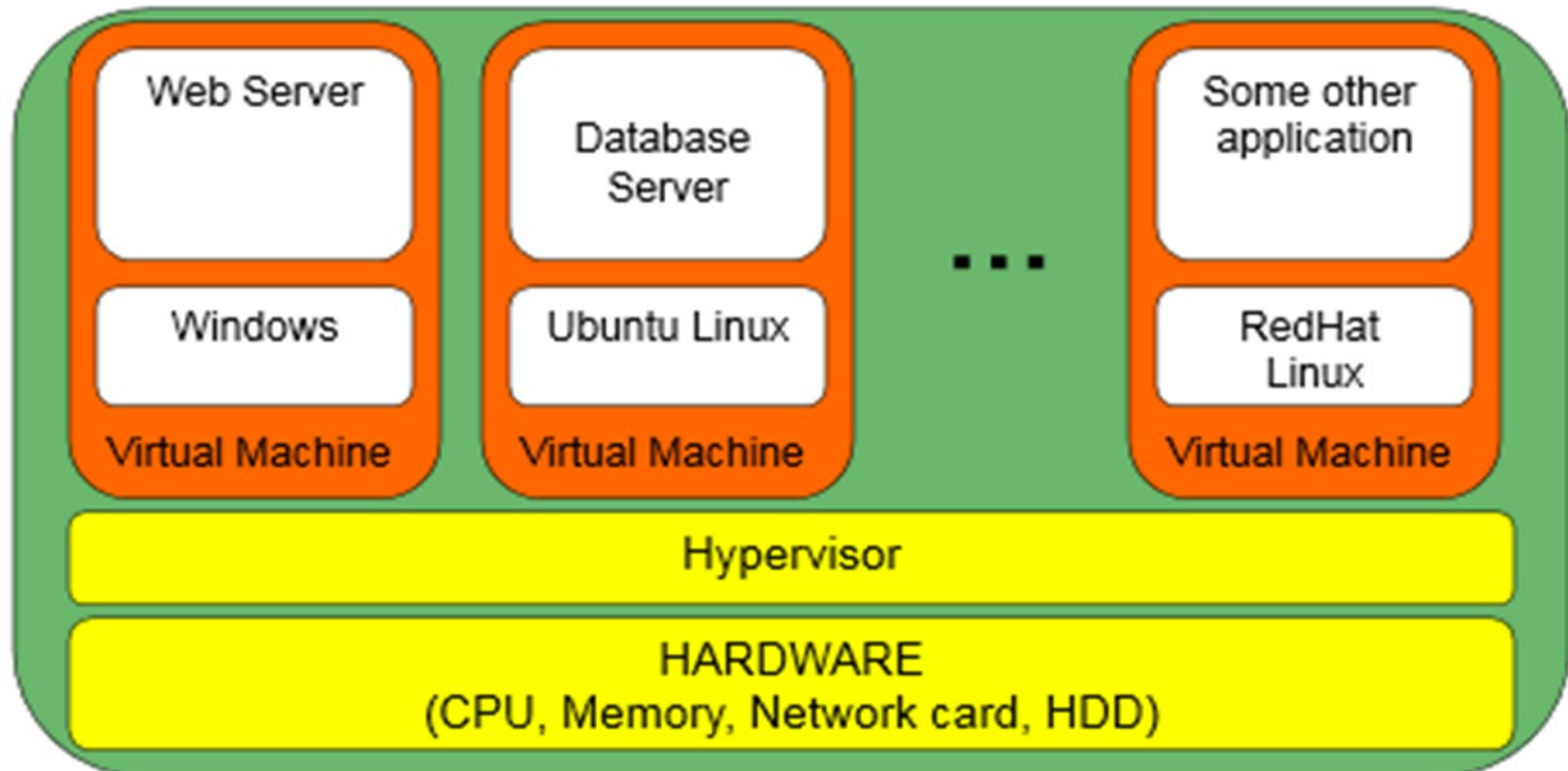
24

- Cloud = Virtualization + Data Center
- Over-the-Internet provisioning of dynamically scalable, virtualized resources
 - ▣ Computing/storage resources, computing platform and middleware, services (IaaS, PaaS, SaaS)
- Users do not need expertise in resource management
 - ▣ Hardware maintenance, system configurations, software upgrades, information updates, etc.
 - ▣ Can focus on the high level problems
- Charged by use, like other utilities

Infrastructure as a Service (IaaS)

25

- Computing resources such as processing and storage
 - ▣ Amazon EC2, S3 (Simple Storage Service)
- Virtual machine: in the cloud data center



IaaS

26

- IaaS focus on offering utility computing resources
- Requirements of IaaS
 - ▣ Scalability
 - System performance should remain the same (or at least similar) in small scale or large scale
 - ▣ Elasticity
 - Large amount of resource provisioning and deployment should be done in a short period of time, such as several minutes or hours
 - Clients should be able to dynamically increase or decrease the amount of infrastructure resources in need
 - (Client initiated, not auto-scaling)

□ Requirements of IaaS

▣ Availability and reliability




- Clients should not worry about any failures at the service provider side
- Clients should be able to access computation resources any time and their computation should be completed in a reasonable time (failures should be masked)
- Data stored in cloud can be retrieved whenever needed
- Communication capability and capacity within the provider domain should be maintained
 - Some consider private networks to the clients, and in this case the entire communication channel to the client should be assured in its availability and capacity

Example: Amazon Web Services (AWS)





28

- Simply refers to the entire cloud suite because everything offered is treated as a web service



Compute & Networking

-  **Direct Connect**
Dedicated Network Connection to AWS
-  **EC2**
Virtual Servers in the Cloud
-  **Route 53**
Scalable Domain Name System
-  **VPC**
Isolated Cloud Resources









Storage & Content Delivery

-  **CloudFront**
Global Content Delivery Network
-  **Glacier**
Archive Storage in the Cloud
-  **S3**
Scalable Storage in the Cloud
-  **Storage Gateway**
Integrates On-Premises IT Environments with Cloud Storage




Database

-  **DynamoDB**
Predictable and Scalable NoSQL Data Store
-  **ElastiCache**
In-Memory Cache







Deployment & Management

-  **CloudFormation**
Templated AWS Resource Creation
-  **CloudTrail**
User Activity and Change Tracking
-  **CloudWatch**
Resource and Application Monitoring
-  **Directory Service**
Managed Directories in the Cloud
-  **Elastic Beanstalk**
AWS Application Container
-  **IAM**
Secure AWS Access Control
-  **OpsWorks**
DevOps Application Management Service
-  **Trusted Advisor**
AWS Cloud Optimization Expert

Analytics

-  **Data Pipeline**
Orchestration for Data-Driven Workflows
-  **Elastic MapReduce**
Managed Hadoop Framework
-  **Kinesis**
Real-time Processing of Streaming Big Data

App Services

-  **AppStream**
Low Latency Application Streaming
-  **CloudSearch**
Managed Search Service
-  **Elastic Transcoder**
Easy-to-use Scalable Media
-  **SES**
Email Sending Service
-  **SQS**
Message Queue Service
-  **SWF**
Workflow Service for Coordinating Components

Applications

-  **WorkSpaces**
Desktops in the Cloud
-  **Zocalo**
Secure Enterprise Storage

- Services that can be managed (created, monitored, terminated), over the Web
- Compute services
 - ▣ Elastic Compute Cloud (EC2): on-demand virtual machines (instances)
 - ▣ Elastic MapReduce (EMR): automatically starts Hadoop for parallel applications:
 - ▣ Auto scaling: seamless +/- number of EC2 instances
- Monitoring services
 - ▣ CloudWatch: monitor resources such as CPU, disk access, network traffic

□ Auto Scaling

- ▣ Replicate EC2 instance to multiple zones/regions to assure performance and fault tolerance
- ▣ Can be user scaling or automatic scaling

□ Elastic Load Balancing

- ▣ Spread incoming request to multiple EC2 instances
 - A user session can stick to a specific EC2 instance
- ▣ Detect the health of EC2 instances, when an unhealthy one is detected, ELB no longer routes traffic to it

Hands-on Example

31

- Getting started with AWS EC2 and how to launch an EC2 instance in the cloud and configure a service



Services ▾

Edit ▾

Qusay H. Mahmoud ▾

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 from our user community, or the AWS Marketplace; or you can select one of your own AMIs.

Quick Start

⏪ ⏩ 1 to 22

My AMIs

AWS Marketplace

Community AMIs

☐ Free tier only ⓘ



Amazon Linux

Free tier eligible

Amazon Linux AMI 2014.09.1 (HVM) - ami-b5a7ea85

The Amazon Linux AMI is an EBS backed image. It includes the 3.14 kernel, Ruby 2.1, PHP 5.5, PostgreSQL 9.3, Docker 1.2, the AWS command line tools, and repository access to many other packages.

Root device type: ebs Virtualization type: hvm



Red Hat

Free tier eligible

Red Hat Enterprise Linux 7.0 (HVM), SSD Volume Type - ami-99bef1a9

Red Hat Enterprise Linux version 7.0 (HVM), EBS General Purpose (SSD) Volume Type

Root device type: ebs Virtualization type: hvm

Case Study: NY Times

32

Self-Service, Prorated Supercomputing Fun!

By DEREK GOTTFRID NOVEMBER 1, 2007 5:30 PM

As part of [eliminating TimesSelect](#), [The New York Times](#) has decided to make all the public domain articles from 1851–1922 available free of charge. These articles are all in the form of images scanned from the original paper. In fact from 1851–1980, all 11 million articles are available as images in PDF format. To generate a PDF version of the article takes quite a bit of work — each article is actually composed of numerous smaller TIFF images that need to be scaled and glued together in a coherent fashion.

- http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/?_php=true&_type=blogs&scp=1&sq=self%20service%20prorated&st=cse&_r=0

NY Times

33

- They used AWS to create PDF files out of TIFF archives (1851 – 1980)
 - ▣ 100 Amazon EC2 instances running Hadoop
 - ▣ Processed 4TB of TIFF images stored in S3
 - ▣ Produced 11 million finished PDFs
 - ▣ Running time: 24 hours
- Sample article: what a Computer meant in 1892
<http://query.nytimes.com/mem/archive-free/pdf?res=9F07E0D81438E233A25751C0A9639C94639ED7CF>

Case Study: CycleComputing

34

- Cycle
- <http://www.cyclecomputing.com/blog/cyclecloud-50000-core-utility-supercomputing>



The screenshot shows the top of a CycleComputing website. The header includes the CycleComputing logo with the tagline 'The Leader in Utility HPC Software' and a navigation menu with links for 'PRODUCTS/SOLUTIONS', 'DISCOVERY & INVENTION', 'PARTNERS', and 'NEWS & BLOG'. Below the header is a large banner for a blog post titled 'CycleCloud Achieves Ludicrous Speed! (Utility Supercomputing with 50,000-cores)'. The post is dated 'Thursday, 19 April 2012 09:00', written by 'admin', and has '0 Comments'. The main text of the post begins with an update: 'Update: Since publishing this blog entry, our 50,000 core CycleCloud utility supercomputer has gotten great coverage by BusinessWeek, TheRegister, the NY Times, the Wall Street Journal's CIO Report, Ars Technica, TheVerge, among many others. And now it would run for \$750/hr with the AWS spot pricing as of 6/22/2012! Click here to contact us for more information...'.

CYCLE COMPUTING
The Leader in Utility HPC Software

PRODUCTS/SOLUTIONS ▼ DISCOVERY & INVENTION ▼ PARTNERS ▼ NEWS & BLOG

CycleCloud Achieves Ludicrous Speed! (Utility Supercomputing with 50,000-cores)

🕒 Thursday, 19 April 2012 09:00
✍️ Written by admin
💬 0 Comments

Update: Since publishing this blog entry, our 50,000 core [CycleCloud utility supercomputer](#) has gotten great coverage by [BusinessWeek](#), [TheRegister](#), the [NY Times](#), the [Wall Street Journal's CIO Report](#), [Ars Technica](#), [TheVerge](#), among many others. And now it would run for \$750/hr with the [AWS spot pricing](#) as of 6/22/2012! [Click here to contact us for more information...](#)

Case Study: CycleComputing

35

- Built a 50,000 core supercomputer called Naga using Amazon cloud infrastructure:
 - ▣ 6,742 instances
 - ▣ 51,132 cores
 - ▣ 58.78TB memory
 - ▣ Ran a Chemistry computational intensive job at \$4,828.85/hr
 - ▣ They estimate the job used over 20 million \$ in infrastructure

Teaching Cloud Computing

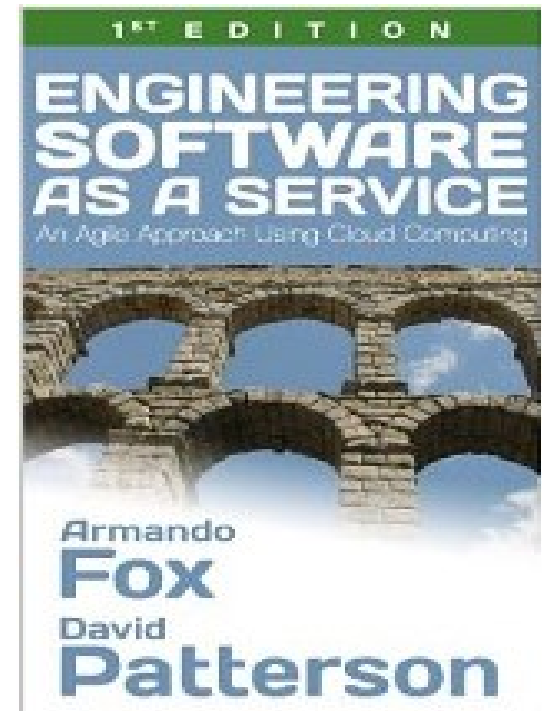
36

- Undergraduate and grad courses on “Distributed Systems” at UoGuelph
 - ▣ Using Amazon EC2 in the lab (for undergrads)
 - ▣ One assignment
 - ▣ Project (use of AWS is optional)
 - Storming the cloud (Dos in the Google App Engine)
- ▣ AWS Education Grant: \$100/student
 - Credit card issues

Teaching Cloud Computing

37

- In my department at UOIT
 - ▣ Assigned a faculty member to teach a special topics course “Software Engineering in the Cloud”
 - ▣ Focuses on building SaaS apps
 - ▣ Recommended book
- ▣ Should cover IaaS and PaaS, and engineering secure SaaS apps



Teaching Cloud Computing

38

□ Graduate course “Cloud Computing” in Spring 2014

The objective of this course is to expose students to the state of the art in cloud computing. Students will learn about issues relevant to the design, implementation and operation of cloud computing infrastructure, platforms, and services. Topics include data centers, virtualization, storage, big data, cloud programming models, services and resource management, and security, privacy and trust issues. Programming assignments will provide students with hands-on experience using widely used cloud environments. In addition, students will learn about systems research through readings and presentations, and a research & development project.

□ Course on Blackboard

Teaching with Cloud Computing

39

- Ok, I am convinced...where should I start?
 - ▣ Develop a new course on cloud computing (& big data)
- No space in the curriculum?
 - ▣ Integrate cloud computing (& big data) into existing courses:
 - Software Engineering
 - Security (hacking on VMs)
 - Databases
 - Networking
 - ...others



Education and Research Grants

40

- Amazon Web Services in Education Grants
 - ▣ <http://aws.amazon.com/grants>

- Google App Engine Education Awards
 - ▣ https://research.google.com/university/relations/appengine/education_awards.html

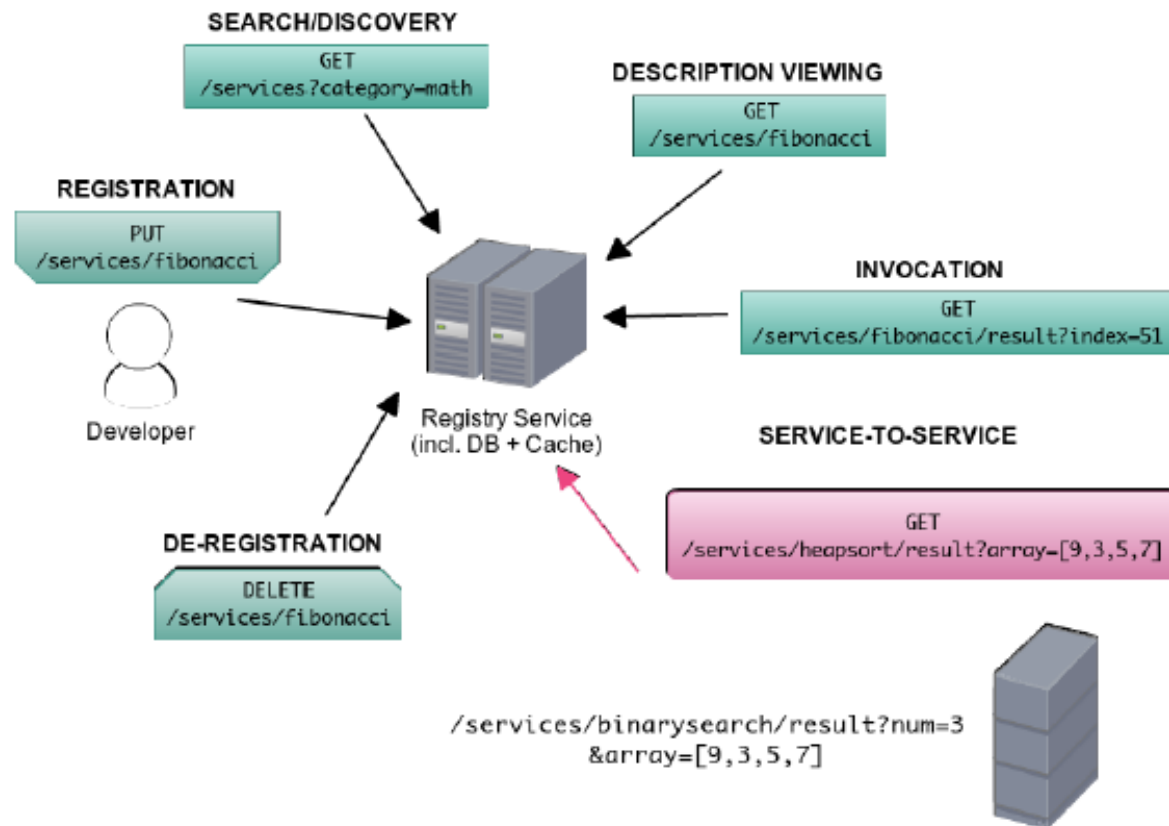
- Microsoft Azure for Research
 - ▣ <http://research.microsoft.com/en-us/projects/azure>

- Others...

Cloud Computing for Research

41

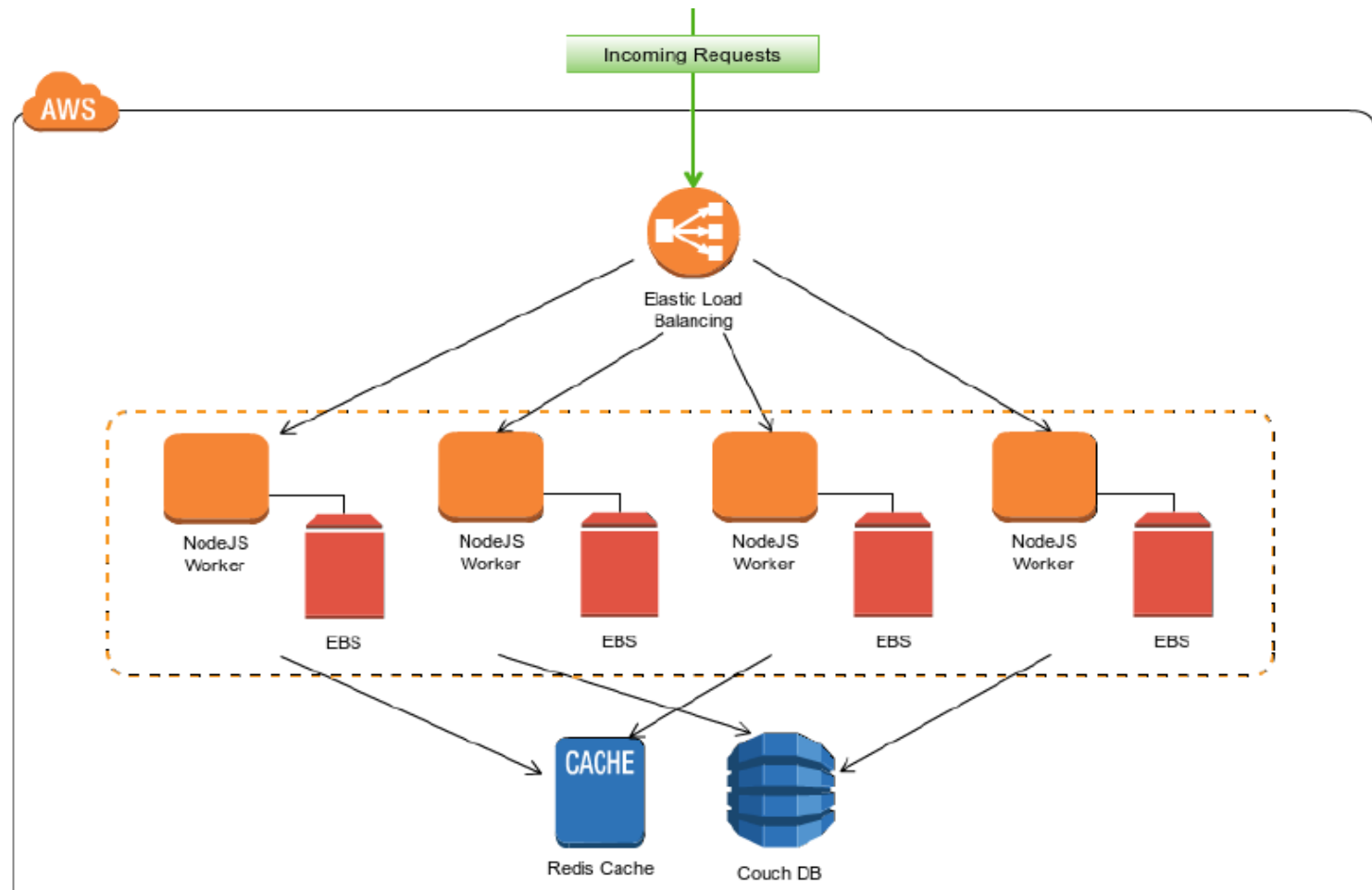
- A platform for provisioning community-contributed web services (with Daniel Vijayakumar)



Cloud Computing for Research

42

□ Implementation of the platform in AWS



Research Opportunities

43

- Cloud computing (and SaaS) adoption in UAE

- Or your institution

- Ankabut.ae



- EBTIC.org : Etisalat BT Innovation Centre

- Others...you may know about...

- Challenges:

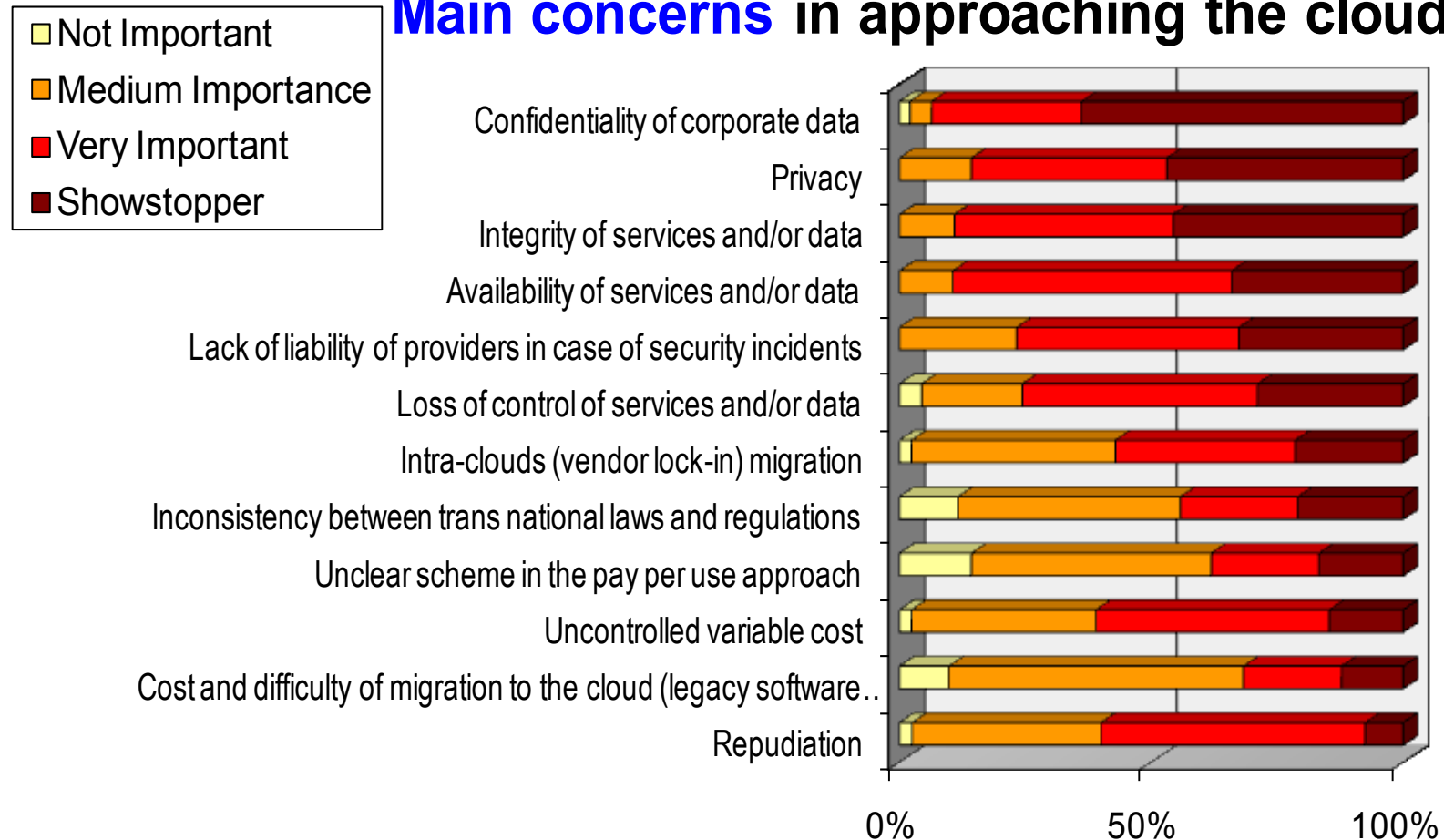
- Security, privacy, trust
 - Portability between providers (standards?)
 - Many others...

Main Concerns of Cloud Computing

44

□ Results from 2009 survey by ENISA

Main concerns in approaching the cloud

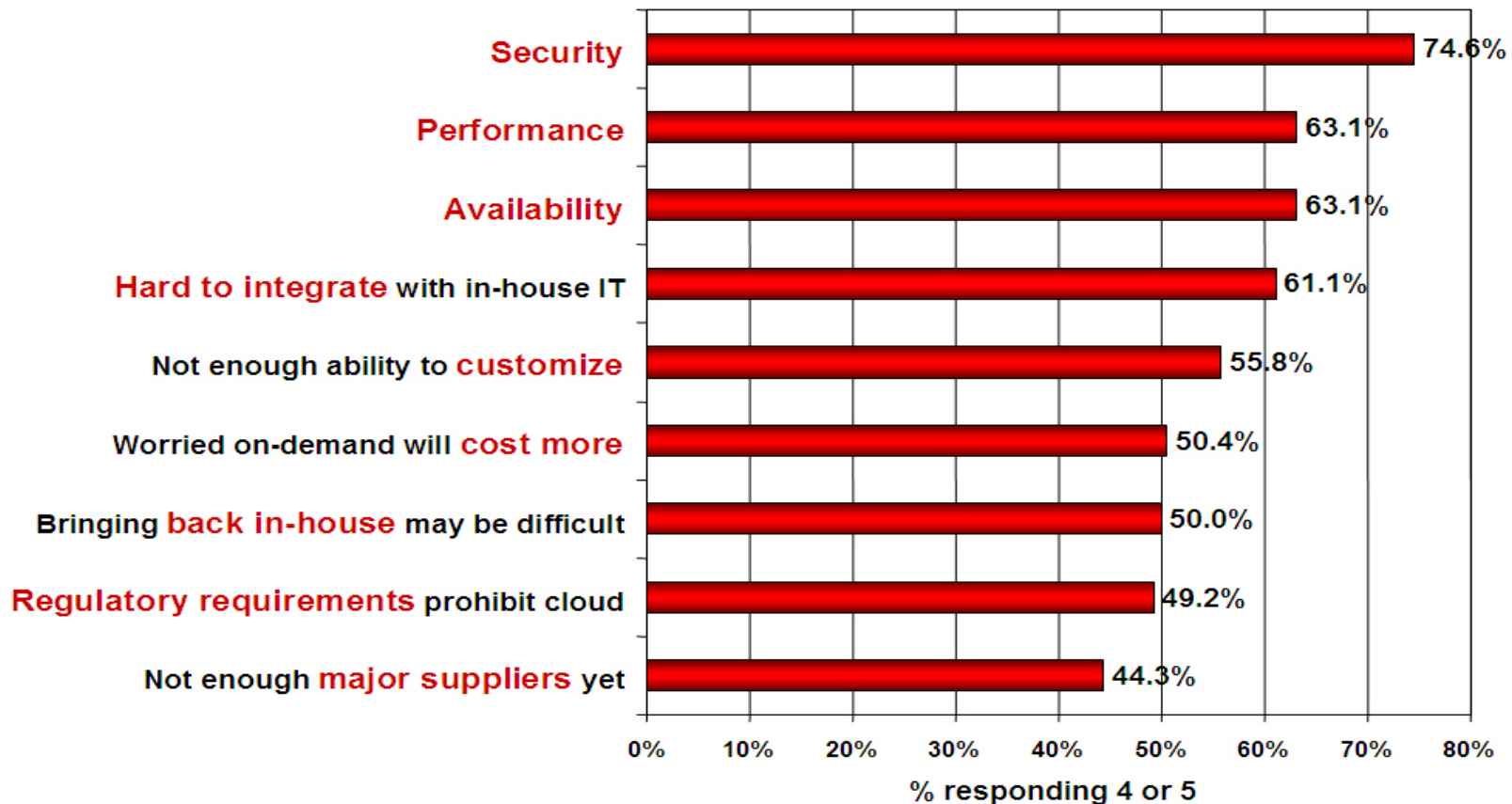


Survey: Security is the major issue

45

Q: Rate the challenges/issues ascribed to the 'cloud'/on-demand model

(1=not significant, 5=very significant)



General Security Challenges

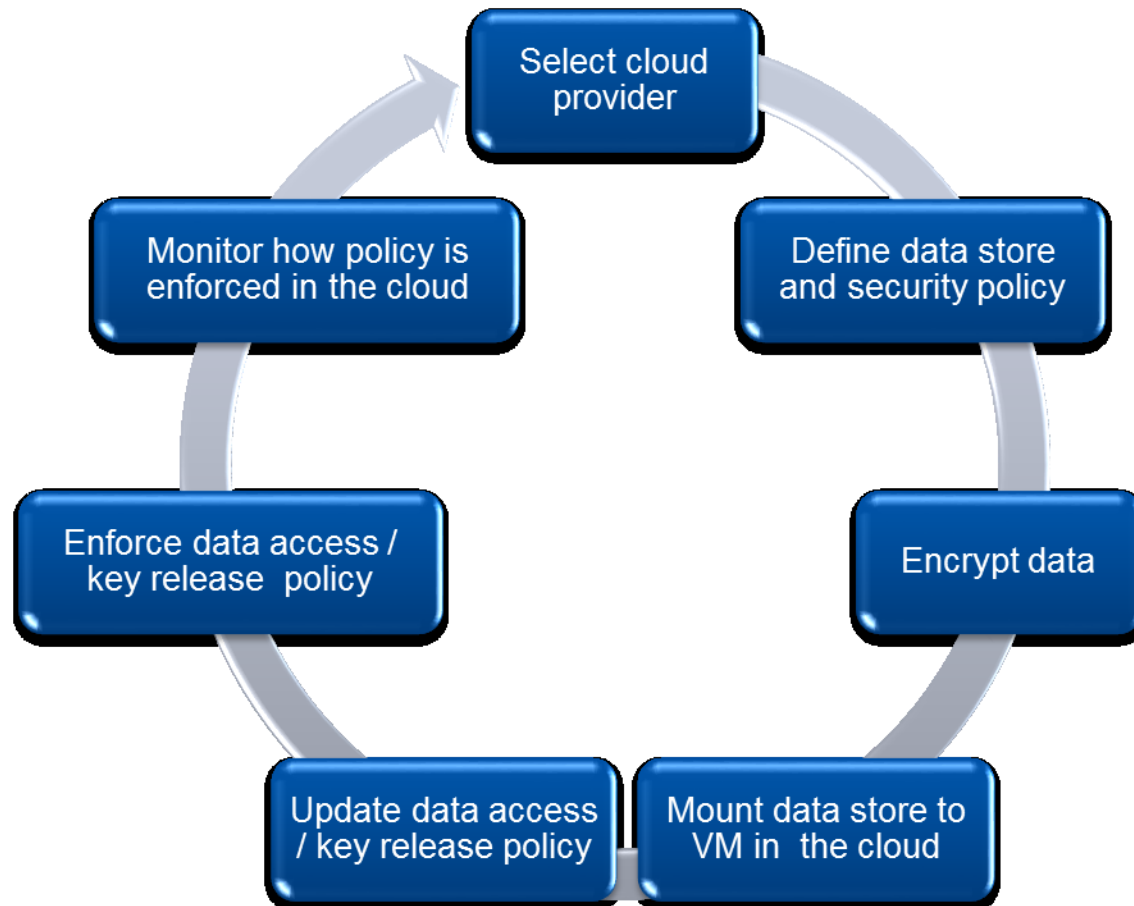
46

- ❑ Insecure interfaces & API's
- ❑ Malicious insiders
- ❑ Shared technology issues
- ❑ Data loss or leakage
- ❑ Loss of physical control
- ❑ Account or service hijacking
- ❑ Trusting the vendor's security model
- ❑ Obtaining support for investigations
- ❑ Others...



Towards a Comprehensive Solution

47



48

Big Data

Big Data

49

- Where does it come from?

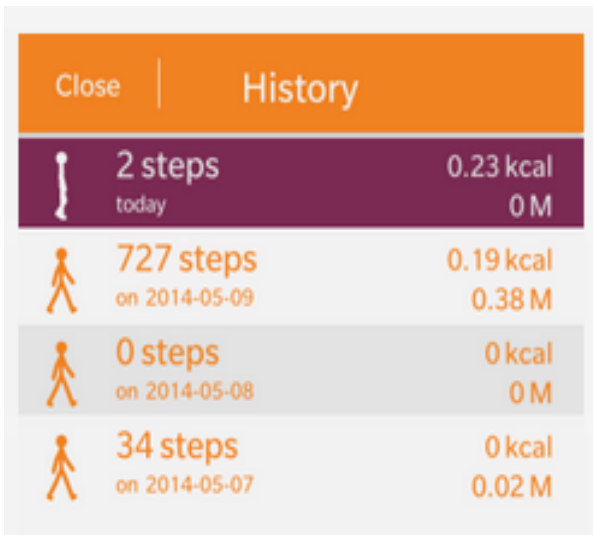
From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days...and the pace is accelerating.

**Eric Schmidt,
*Executive Chairman, Google***





Where does it come from?

50

- Datafication of the world – generating data at freighting rates
 - ▣ Activity data: when you search the web, when you shop (credit card transactions, etc), when you read an ebook, even when you walk!
 - ▣ Conversations data: Twitter, Facebook
 - ▣ Sensor data
 - ▣ IoT data



A screenshot of a mobile application interface showing a 'History' tab. The interface has an orange header with 'Close' and 'History' buttons. Below the header, there is a list of activity records. Each record includes a stick figure icon, the number of steps, the date, and the calories burned. The first record is highlighted in purple and shows '2 steps today' and '0.23 kcal 0 M'. The other records are in white and show '727 steps on 2014-05-09' (0.19 kcal, 0.38 M), '0 steps on 2014-05-08' (0 kcal, 0 M), and '34 steps on 2014-05-07' (0 kcal, 0.02 M).

	Close	History
	2 steps today	0.23 kcal 0 M
	727 steps on 2014-05-09	0.19 kcal 0.38 M
	0 steps on 2014-05-08	0 kcal 0 M
	34 steps on 2014-05-07	0 kcal 0.02 M

How much data does an airplane generate in one trip?

51

- Etihad Airways uses Big Data to reach its destination

How Etihad Airways Uses Big Data to Reach Its Destination



Posted August 6, 2013



comments



- <http://smartdatacollective.com/bigdatastartups/137741/how-etihad-airways-uses-big-data-reach-its-destination>
- Every sensor, every battery, every video screen watched by a passenger...

Governments Big Data

52

- US Government: <http://www.data.gov>
- City open data (many available online)
 - ▣ (e.g. Toronto Open Data)
- UAE Open Data



government.ae
The official portal of the United Arab Emirates

Home

Contact us

Sitemap

Help

-A

+A



About UAE

eServices

MGovernment

Open Data

Contact Gove



The Digital Universe

53

□ The Digital Universe



The Economist, Feb 25, 2010

IN 2010 THE DIGITAL UNIVERSE WAS
1.2 ZETTABYTES

IN A DECADE THE DIGITAL UNIVERSE WILL BE
35 ZETTABYTES

90% OF THE DIGITAL UNIVERSE IS
UNSTRUCTURED

IN 2011 THE DIGITAL UNIVERSE IS
300 QUADRILLION FILES

Data-Driven Apps

54

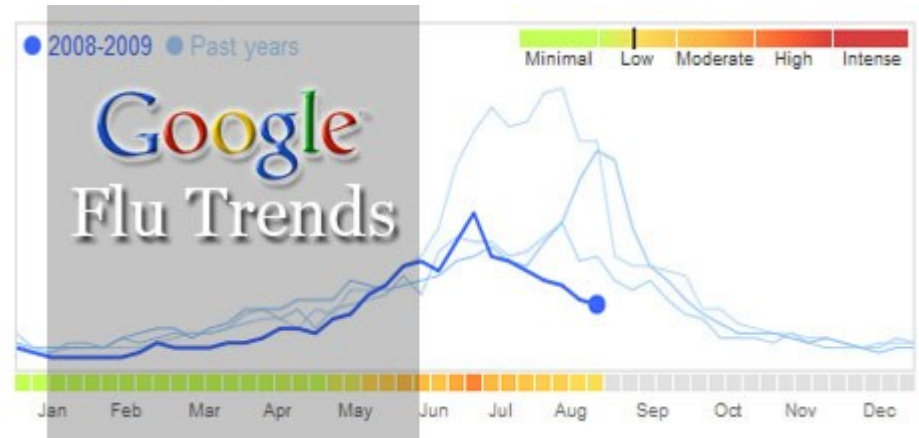
- Google Flu / Correlate / Trends
- NextBus

Tracked vehicles for route Campus Shuttle arriving in:

3 minutes
7 minutes
18 minutes



Stop Number: 102
Phone: 717-781-8925
SMS: 41411 "yorkc 102"



- Data analytics examples:
- Wefeelfine
- Google correlate
- Google flu

Toronto Transit Commission

Choose another transit system...

Select your route/direction/stop to obtain GPS-based arrival times:

Route: 29-Dufferin

Direction: South - 29a Dufferin towards Exhibition (Dufferin Gate)

Stop: Dufferin Street At Eversfield Road Farside

This system is currently in Beta testing

Tracked vehicles for route 29-Dufferin arriving in:

2 minutes

South - 29d Dufferin towards Exhibition (Princes' Gates)

2 minutes

South - 29 Dufferin towards Exhibition (Dufferin Gate)

17 minutes

South - 29 Dufferin towards Exhibition (Dufferin Gate)

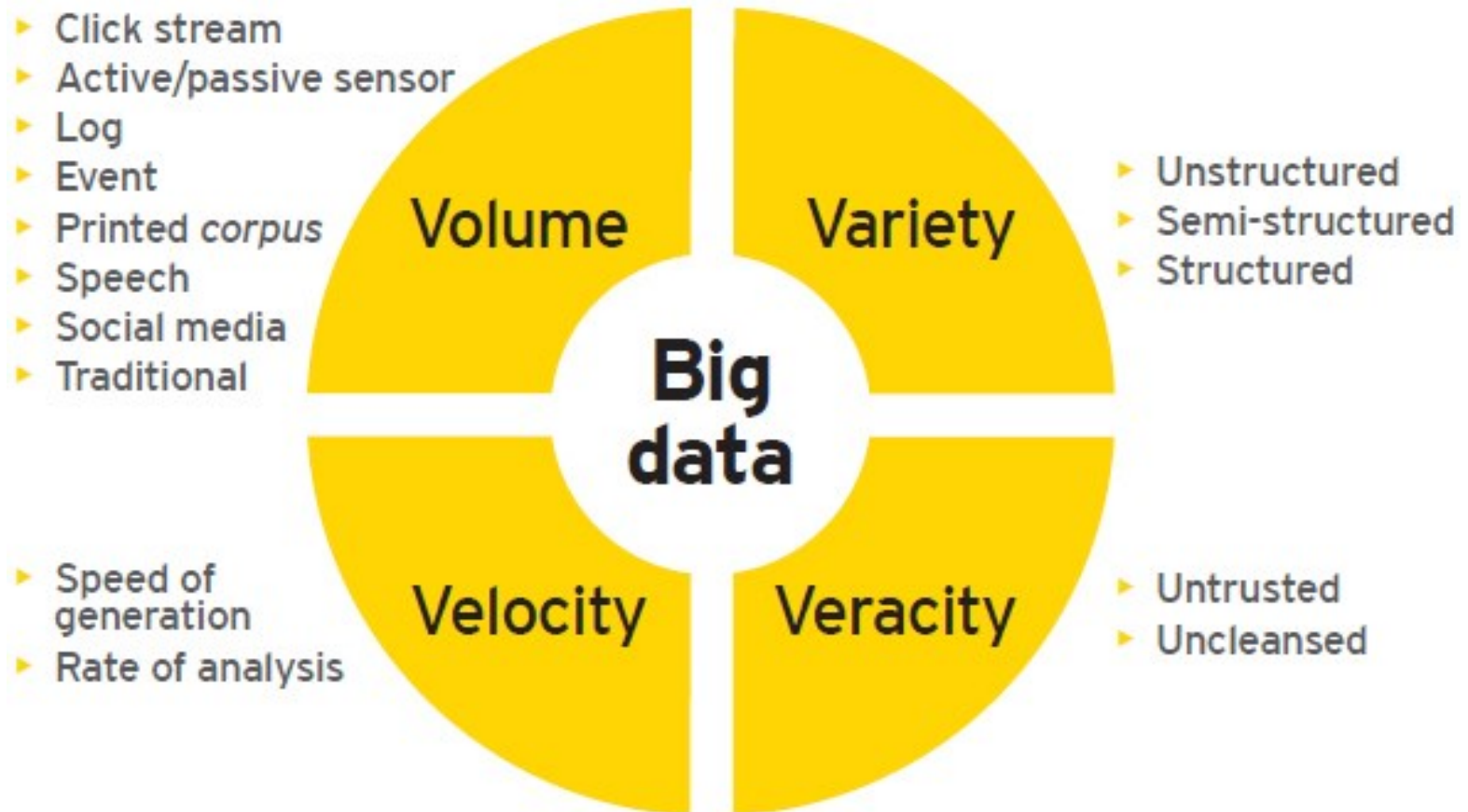
Valid as of 10:30 AM Tuesday, May 10

Go to page that can be bookmarked ?

Big Data

55

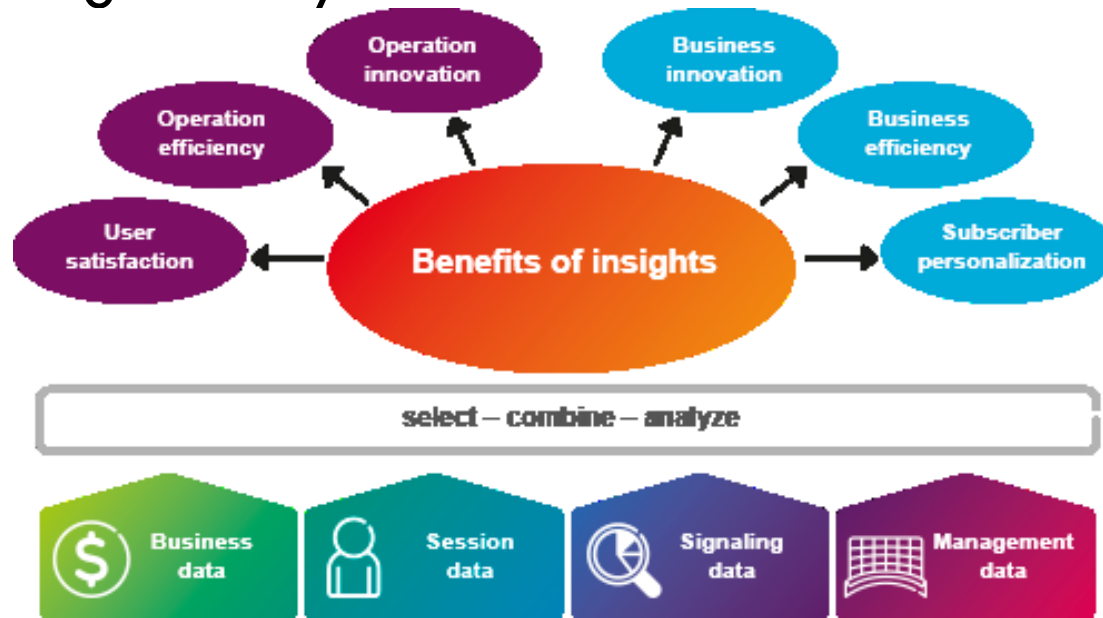
□ The four V's of big data



Big Data Analytics

56

- Turning big data into value
 - ▣ Better understand and target customers
 - ▣ Improved business processes
 - ▣ Improving health
 - ▣ Improving security and law enforcement
 - ▣ ...etc



Large Scale Data processing

57

□ Big Data Technologies

- Google → MapReduce, Sawzall, BigQuery

- Yahoo → Pig Latin

- Microsoft → Dryad, DryadLINQ

 - Dropped Dyrad, focuses on Hadoop

- Berkeley Data Analytics Stack



- Apache Hadoop is an open source implementation of MapReduce

 - Hive

 - Pig (open source of Yahoo Pig Latin)

What is Hadoop?



58

- A scalable fault-tolerant distributed system for data storage and processing
- At the core, there are two main components:
 - ▣ Hadoop Distributed File System (HDFS): high-bandwidth clustered distributed file system optimized for large files
 - ▣ MapReduce: programming model for processing sets of data; mapping inputs to outputs and reducing the output of multiple Mappers to a single (or a few) answers
- Operating on unstructured and structured data
- <http://hadoop.apache.org>

Why Hadoop?

59

Big Data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to address business and technology trends that are disrupting traditional data management and processing.

Enterprises can gain a competitive advantage by being early adopters of big data analytics.

Gartner®

Hadoop Adoption

60

2007

YAHOO!



lost.fm

2008

Google
ImageShack
ablegrape
Cascading

IBM
facebook

ENORMO
Every property. Everywhere.
A9
krugle
rackspace
HOSTING

THE UNIVERSITY OF EDINBURGH
rackspace
HOSTING

Lookery
Control freaks welcome

The New York Times
Joost

Zvents
Discover Things To Do
FORMATION SCIENCES INSTITUTE

#News Corporation

Cornell University
Computing and Information Science
Visible MEASURES

LOTAME
Locality, Target, & Message with Social Media
NetSeer

parc
Palo Alto Research Center
SECURITY ENHANCED
BRANCH NAME SYSTEM
veoh

2009

AOL
cloudera

deepdyve
cooliris

eyealike
TEXTMAP
THE CITY'S SEARCH ENGINE

PG College of Technology
iterend

tailsweep
hulu

RapLeaf
USCMS

Ning
quxntcast

amazon
web services
pressflip

detikSearch
WorldLingo

Systems@ETH Zürich

VK SOLUTIONS
Global Solutions Provider
TARAGANA
Innovation + Quality + Simplicity

HOSTING HABITAT
HOLA
by solutions

Terrier
adknowledge

stampede
beta

2010

SAMSUNG
rubicon

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY
VISIBLE TECHNOLOGIES

APOLLO GROUP
ADSDAQ

rackspace
HOSTING
RapLeaf

wordnik
At the words
MOBILGEN
Mobile Search & Analytics

comSCORE
trulia
real estate search

Accela COMMUNICATIONS
Forward3D

LinkedIn
Microsoft

Infochimps
Find the world's data
Pharm 2PhorK

ADMELD
gumgum
BrainPad

Pronux
The Datagraph Blog

NETFLIX
mobileanalytics.tv

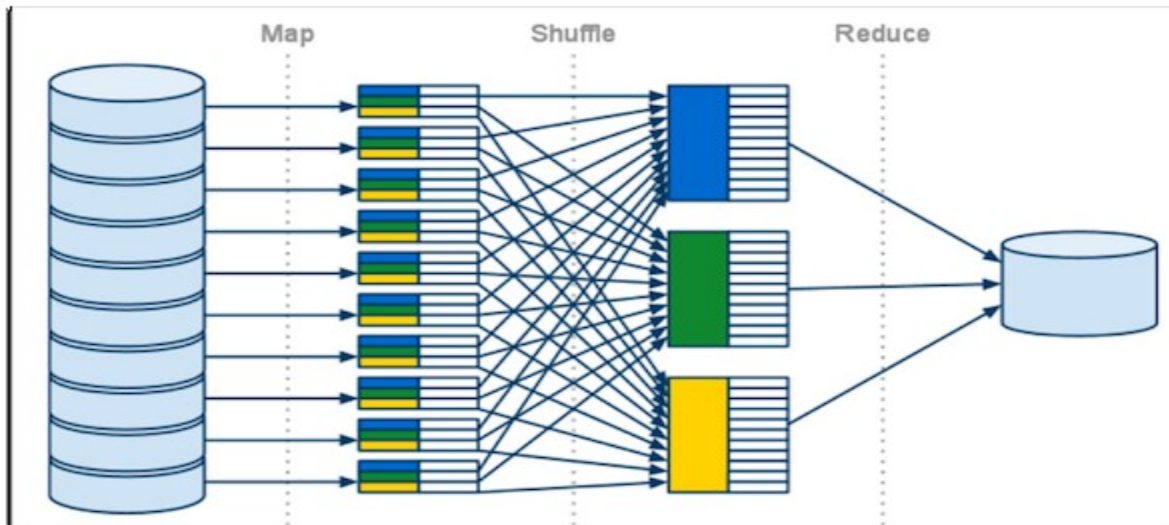
markt24.de
twitter

media6degrees
BEEBLER
SLC Security
When Experience Matters...
eBay

Google MapReduce

61

- Hadoop implements MapReduce:
 - ▣ An abstraction that allows programmers to specify computations that can be done in parallel; expressed in 2 functions: map, reduce
 - ▣ A method for distributing a task across multiple nodes
 - ▣ Each node processes data stored on that node



MapReduce Programming Model

62

□ Map

- ▣ Takes an input pair and produces a set of intermediate key/value pairs e.g.,
 - Map: $(key_1, value_1) \rightarrow list(key_2, value_2)$
 - The MapReduce library groups together all intermediate values associated with the same intermediate key

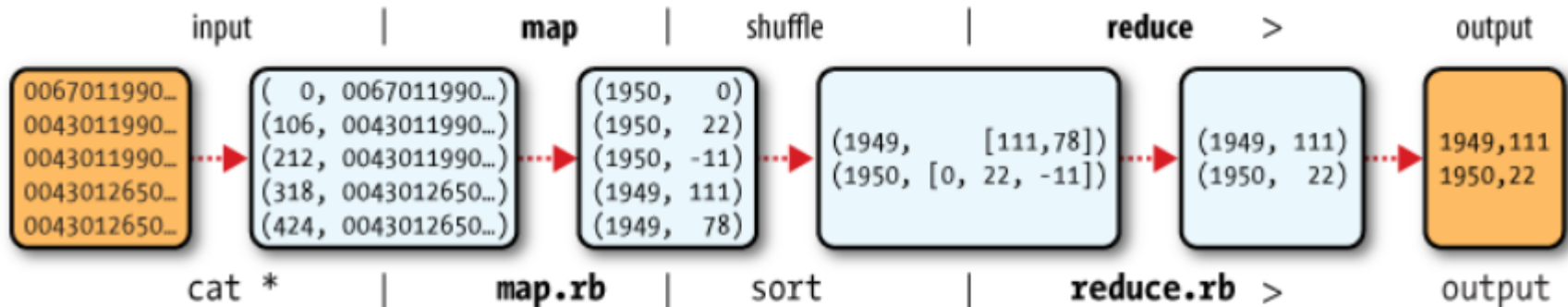
□ Reduce

- ▣ This function accepts an intermediate key and a set of values for that key
 - Reduce: $(key_2, list(key_2, value_2)) \rightarrow value_3$

Example 1

63

- Man/min temperature for the last century?

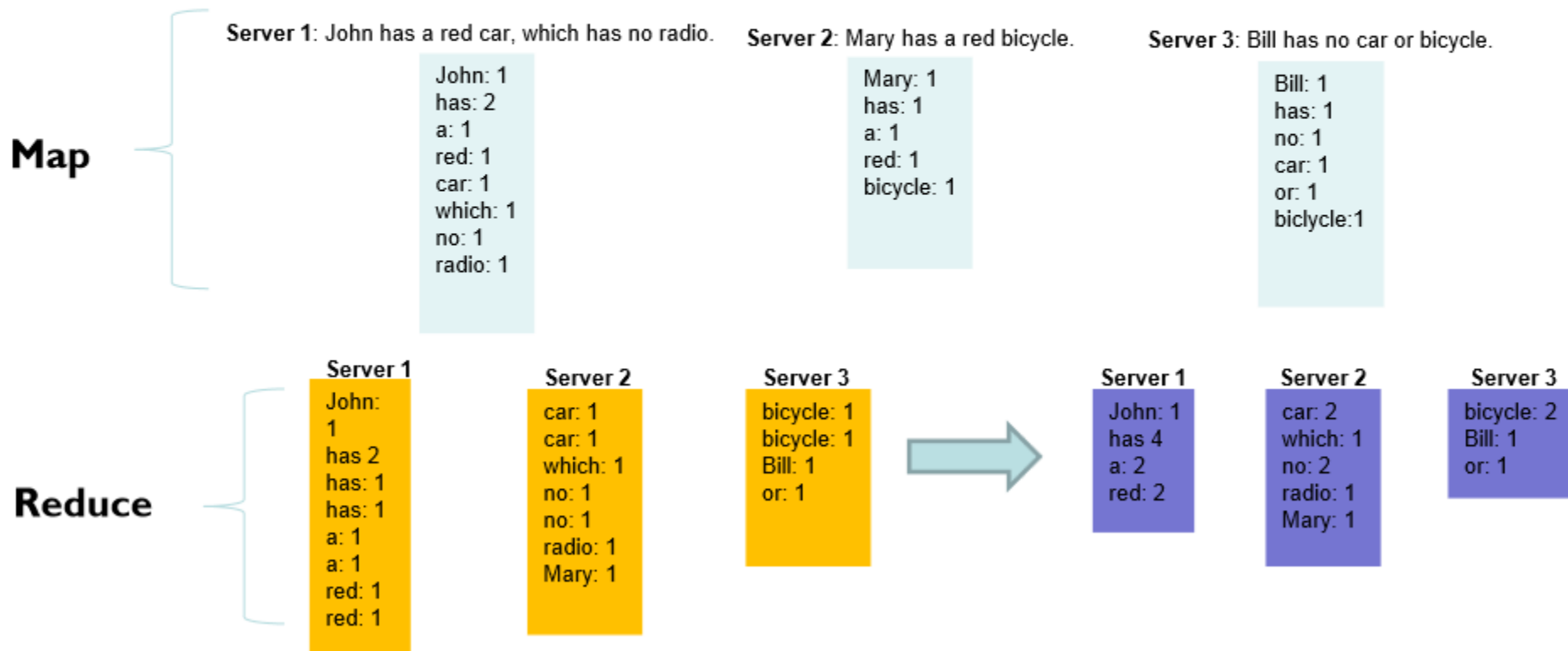


What was the max/min temperature for the last century?

Example e

64

Problem: Count the number of times that each word appears in the following paragraph:
John has a red car, which has no radio. Mary has a red bicycle. Bill has no car or bicycle.



Example 3: detailed

65

- Word frequencies in a document or a file
- Determine the count of each word that appears in a document (or a set of documents)
 - Each file is associated with a document URL
- Map function
 - ▣ Key = document URL
 - ▣ Value = document contents
- Output of map function is (potentially many) key/value pairs
 - ▣ Output (word, “1”) once per word in the document

Example 3

66

- “file.txt”, “to be or not to be”
- Applying the map function will produce:

- “to”, 1
- “be”, 1
- “or”, 1
- “not”, 1
- “to”, 1
- “be”, 1

```
Map(String key, String value):  
  // input_key: document name  
  // input_value: document  
  contents  
for each word w in value:  
  EmitIntermediate(w, "1");
```

Example 3

67

- Pseudo code for the produce function
 - ▣ Sums all counts emitted for a particular word

Reduce(String key, values):

// key: a word, same for input and output

// values: a list of counts

int result = 0;

for each v in values:

 result = result + value;

Emit(result);

Example 3

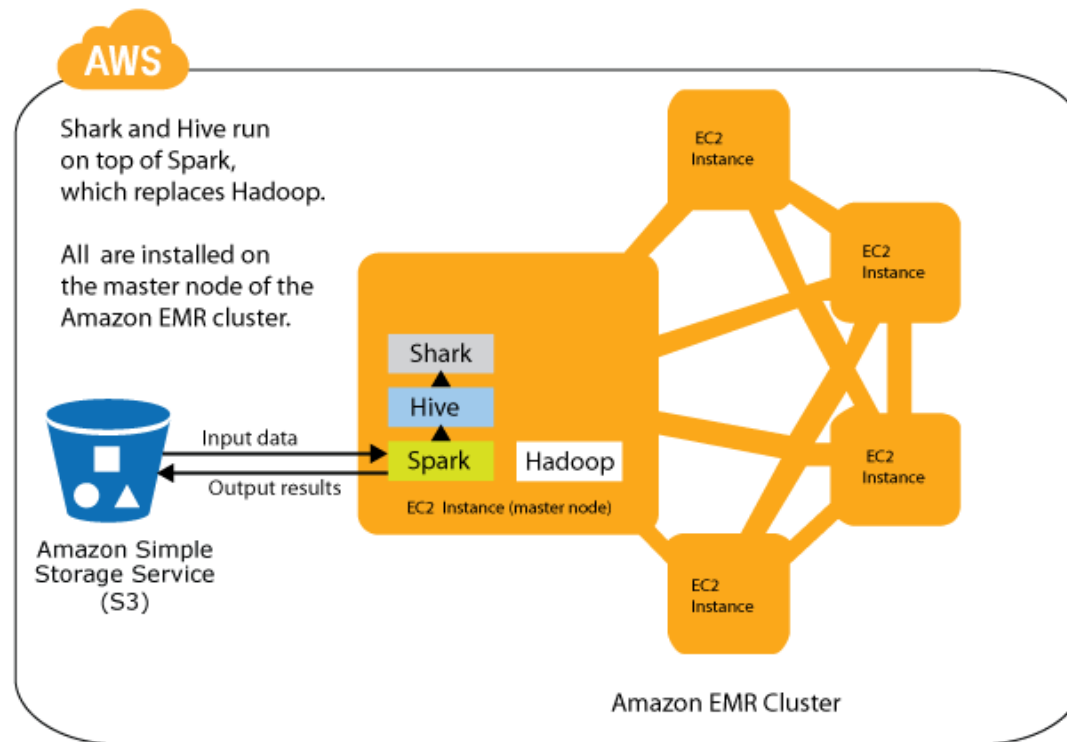
68

- Results
- The MapReduce framework sorts all pairs with the same key
 - ▣ $(be, 1), (be, 1), (not, 1), (or, 1), (to, 1), (to, 1)$
- The pairs are then grouped
 - ▣ $(be, 1, 1), (not, 1), (or, 1), (to, 1, 1)$
- The reduce function combines (sums) the values for a key
 - ▣ Example: Applying reduce to $(be, 1, 1) = 2$

Hands-on Example

69

- Given the limited time we will not go through installing and configuring Hadoop
- We will use Amazon Elastic MapReduce (EMR)



Big Data Security & Privacy Issues

70

□ Discussion

Teaching

71

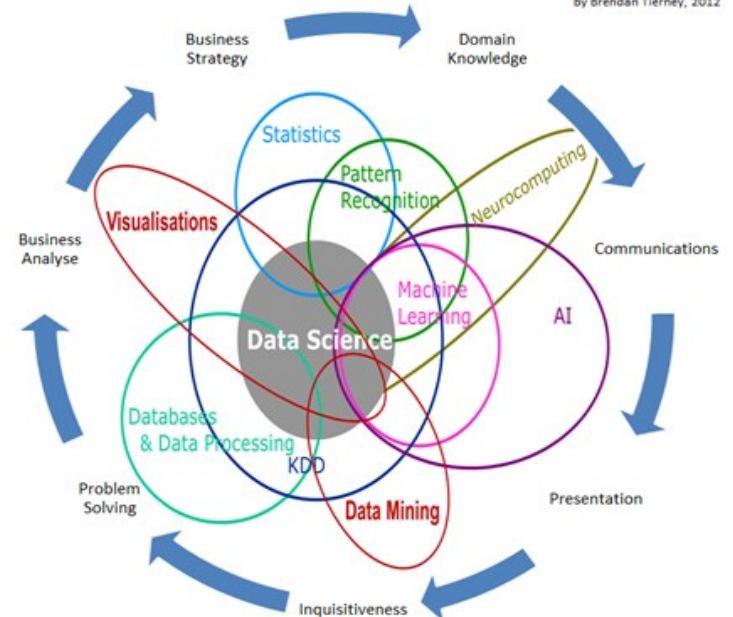
- Grad course on cloud computing
 - ▣ Students install and configure Hadoop in the cloud
 - ▣ Run test applications, and then can start developing

- Datasets:
 - ▣ Project Gutenberg
 - ▣ Teradata University Network
 - ▣ Amazon Public Data Sets

- Thinking of starting a new degree?

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Research

72

- Machine learning
- Data mining
- Statistics

UAE betting big on big data, as CIOs plan analytics investments

***Summary:** As events in the region bring demographic changes, IT chiefs are turning to big data tech to help get more insight from the information they have.*

Source: <http://www.zdnet.com/uae-betting-big-on-big-data-as-cios-plan-analytics-investments-7000026036/>

- Solutions for local markets
 - ▣ Analytic tools for Arabic content

References

73

- Some of the slides have been adapted from:
- <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
- Lew Tucker: Introduction to Cloud Computing for Enterprise Users
- https://media.amazonwebservices.com/AWS_Overview.pdf
- <http://www.utdallas.edu/~ilyen/course/cloud/home.html>
- <http://www.csd.uwo.ca/faculty/hanan/cs843>
- Rob Peglar: Introduction to Analytics and Big Data – Hadoop
- Srijith Nair, Theo Dimitrakos: On the Security of Data Stored in the Cloud
- Google images

Supplementary Online Material

74

□ <http://faculty.uoit.ca/mahmoud/iit2014tutorial.html>

Summary and Discussion

A flavor of cloud computing & big data analytics for teaching and research

Some questions can't be answered by



Qusay.Mahmoud@uoit.ca