UNIVERSITY OF ONTARIO
INSTITUTE OF TECHNOLOGY

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

IIT'14, Nov 7

UAEU University, Al Ain, UAE

Agenda

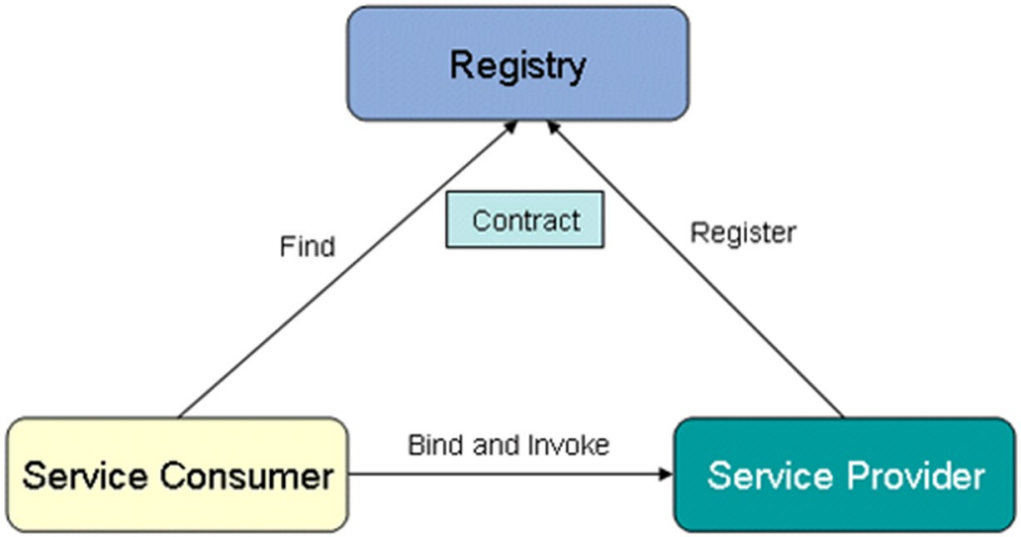
- Cloud Computing
  - What is it? and Why?
  - Service models and cloud providers
  - Hands-on example: using the cloud
  - Security and privacy issues
  - Teaching and Research
- Big Data Analytics
  - What is it Big Data? Where does it come from?
  - Technologies for processing large data sets (analytics)
  - Hands-on example: processing big data
  - Teaching and Research

Learning Outcomes

- 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

- 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

- 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

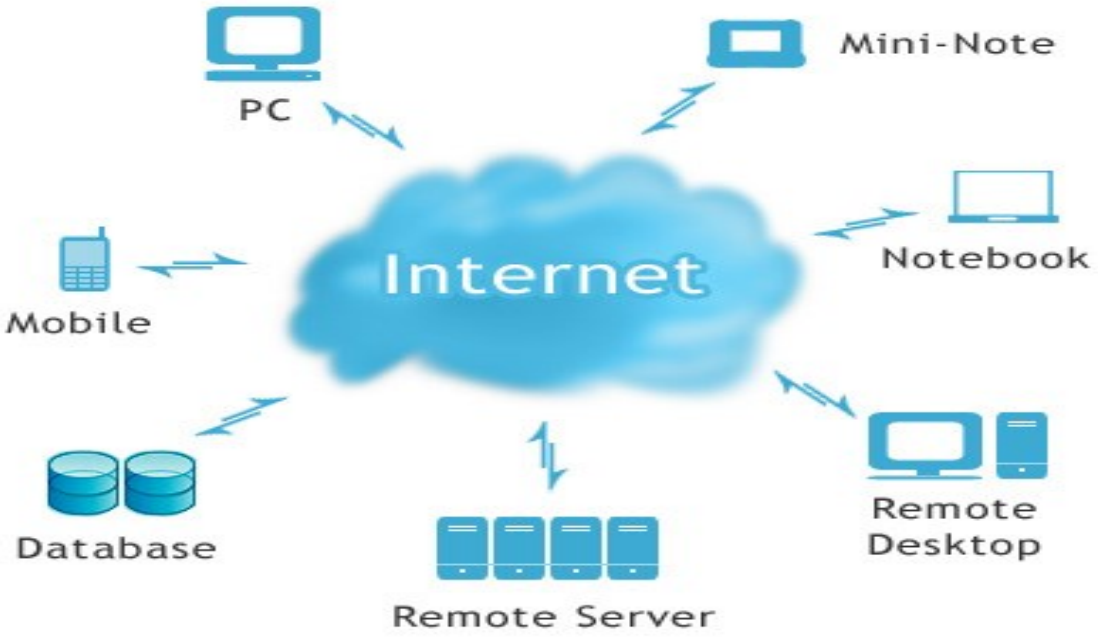
- Benefits
  - Reduced acquisition and maintenance costs (customer)
  - Scalability and QoS is responsibility of service provider (customer)
  - Easier support and maintenance (service provider)
- Downside
  - Security: confidential info visible to others (custome
  - Availability and reliability are harder to guarantee (service provider)

Introduction to Cloud Computing

- 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'?

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



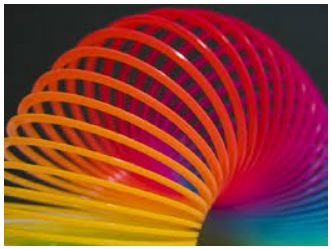
Five Characteristics

- On-demand self-service
  - It's there when you need it

- Broad network access
  - Connectivity options

- Resource pooling
  - Sharing resources (undisclosed location)

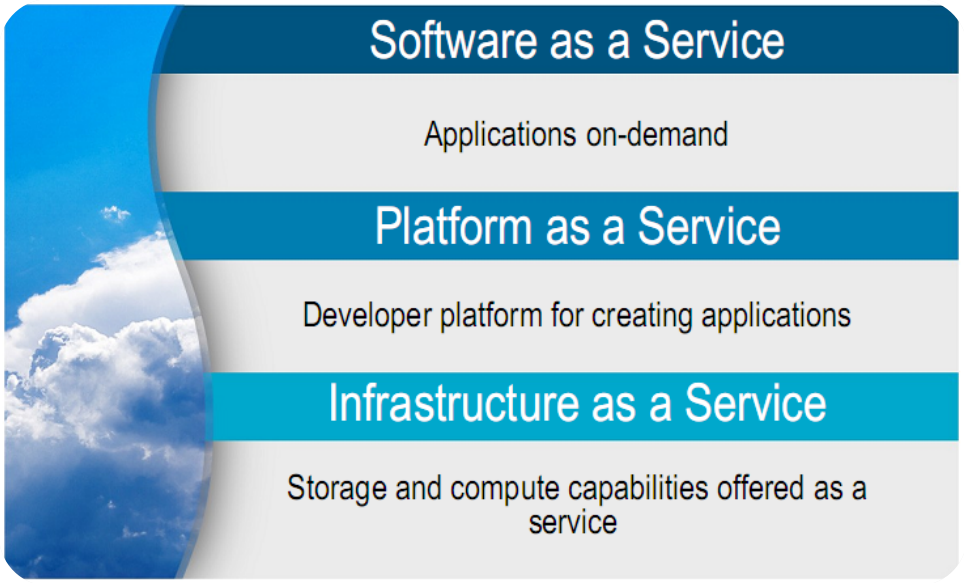
- Rapid elasticity
  - You get what you need

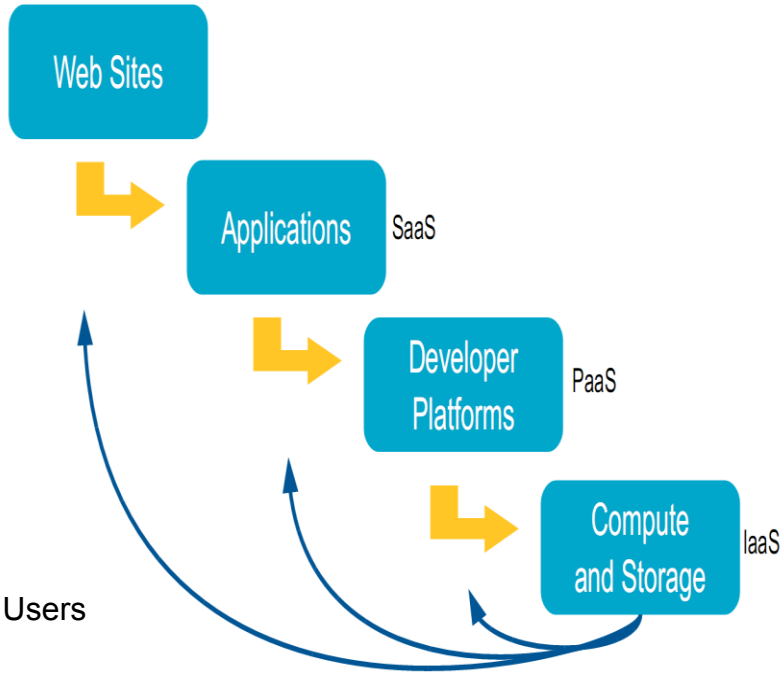


- Measured service
  - You pay for what you use

Three Service Models

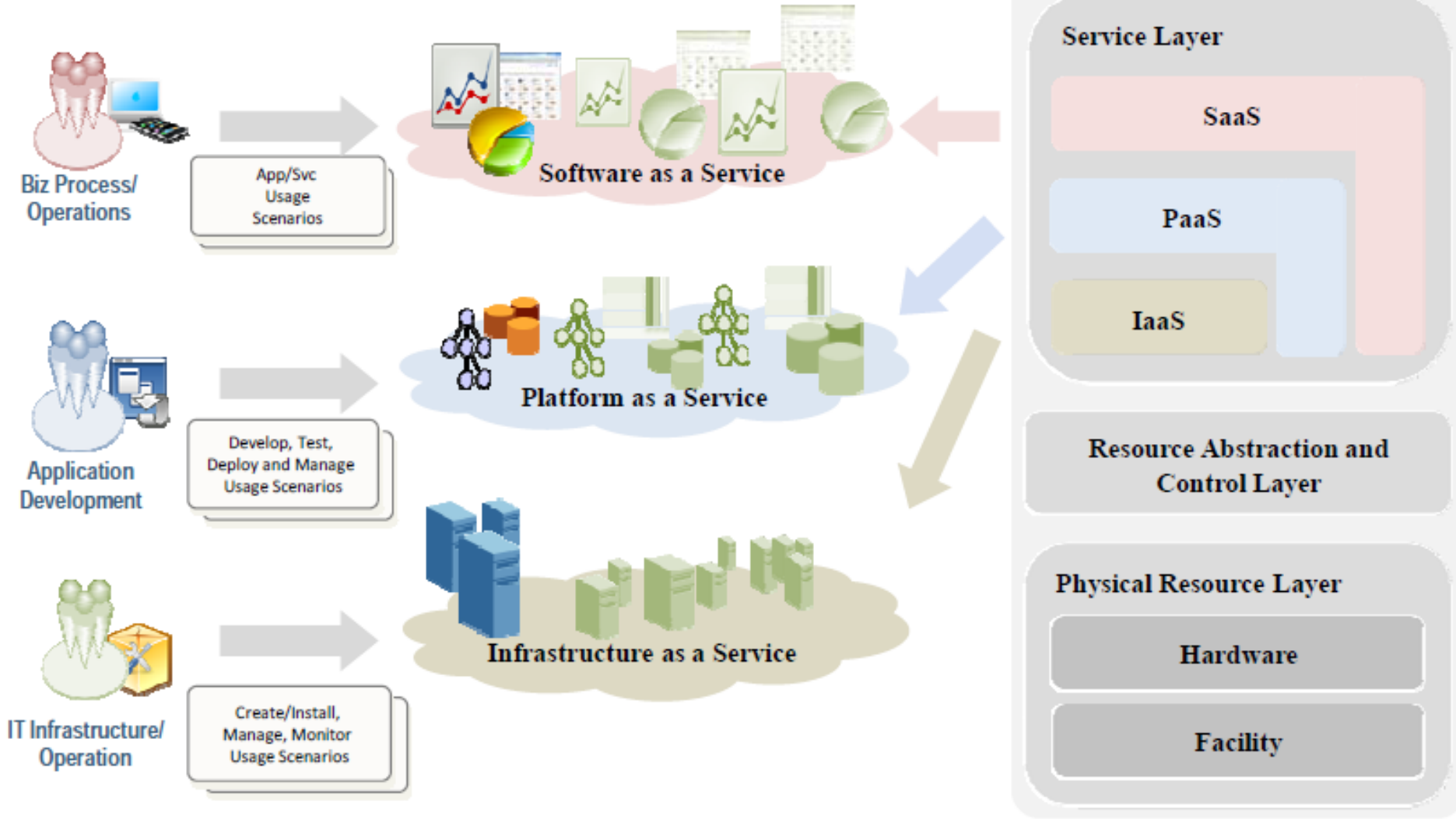
- Software as a service (SaaS)
- Platform as a service (PaaS)
- Infrastructure as a service (IaaS)





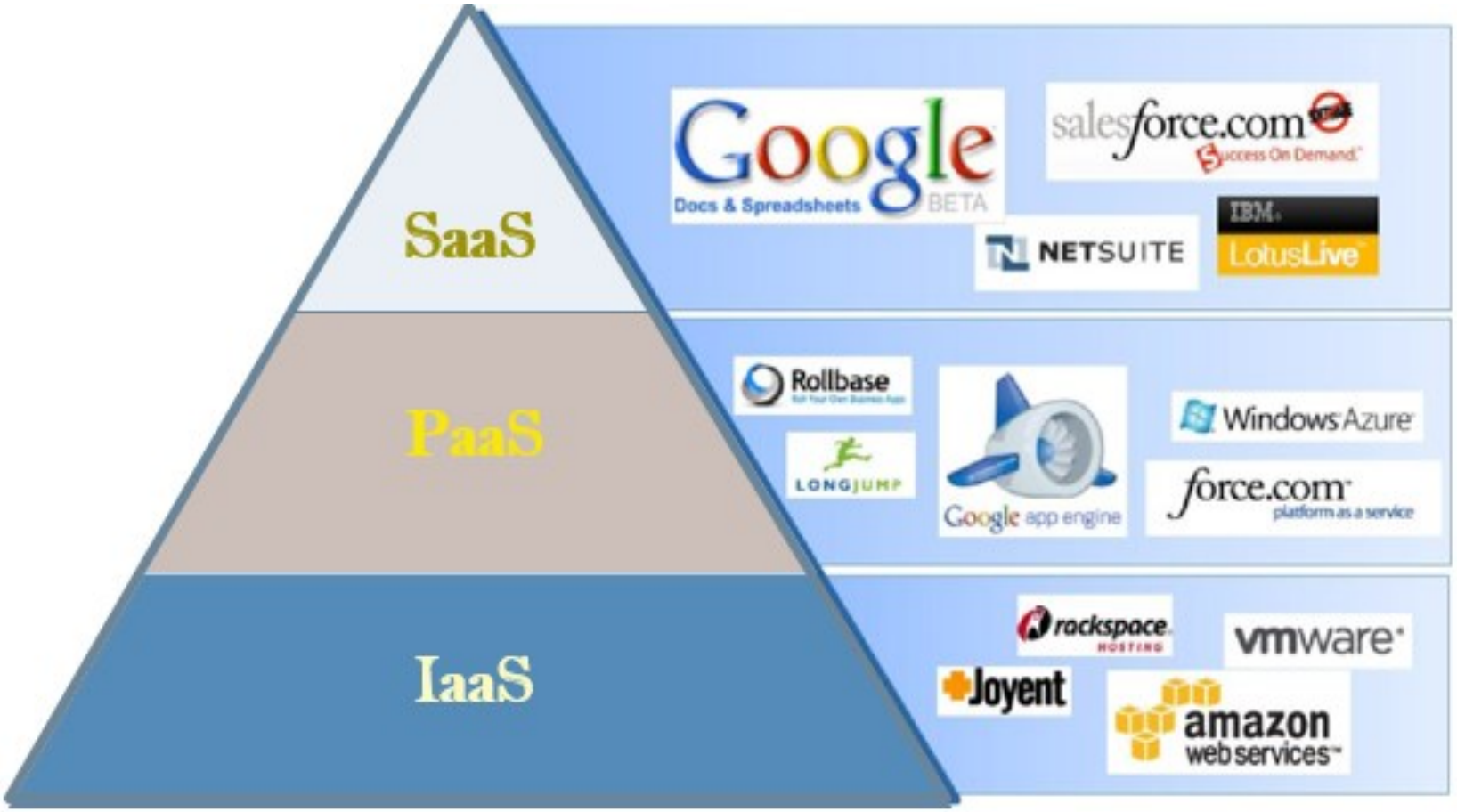
Source: Lew Tucker: Introduction to Cloud Computing for Enterprise Users

Service Orchestration



Cloud Computing Landscape

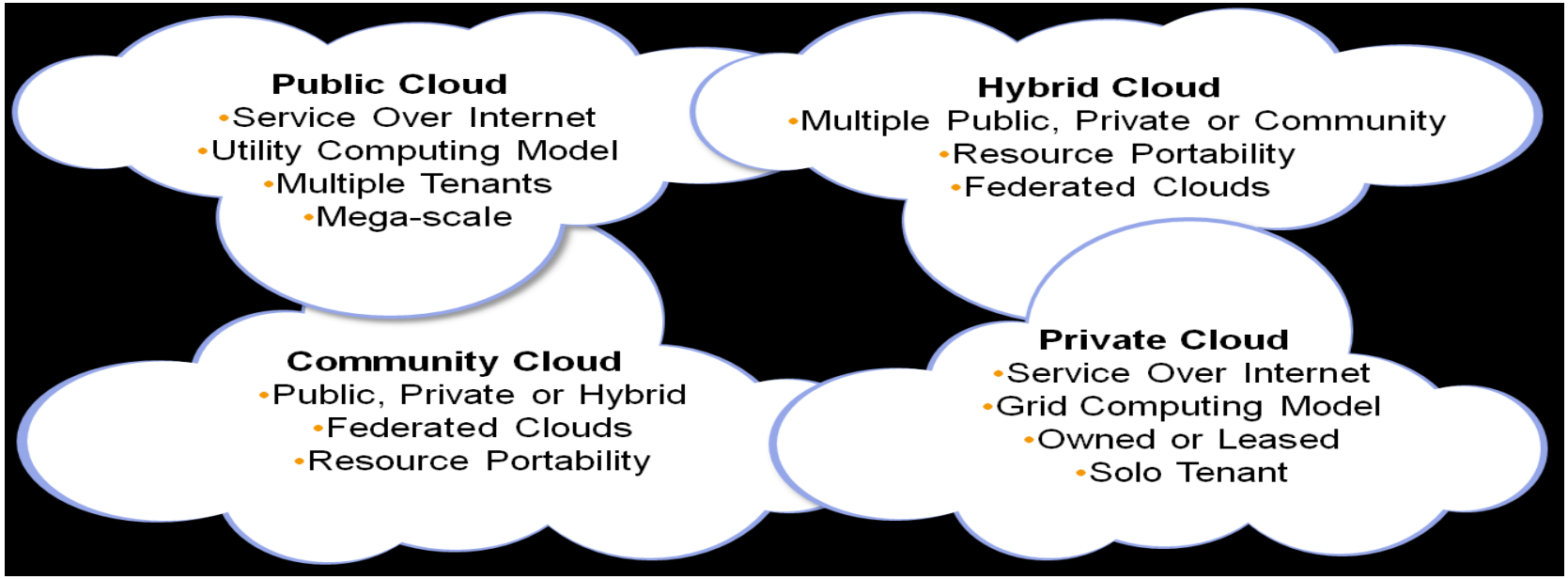
- Some of the cloud providers (or players in this space)



Source: Google images

Four Deployment Models

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



Source: Google images

Utility Computing Business Model

- Cloud computing…the 5th utility?



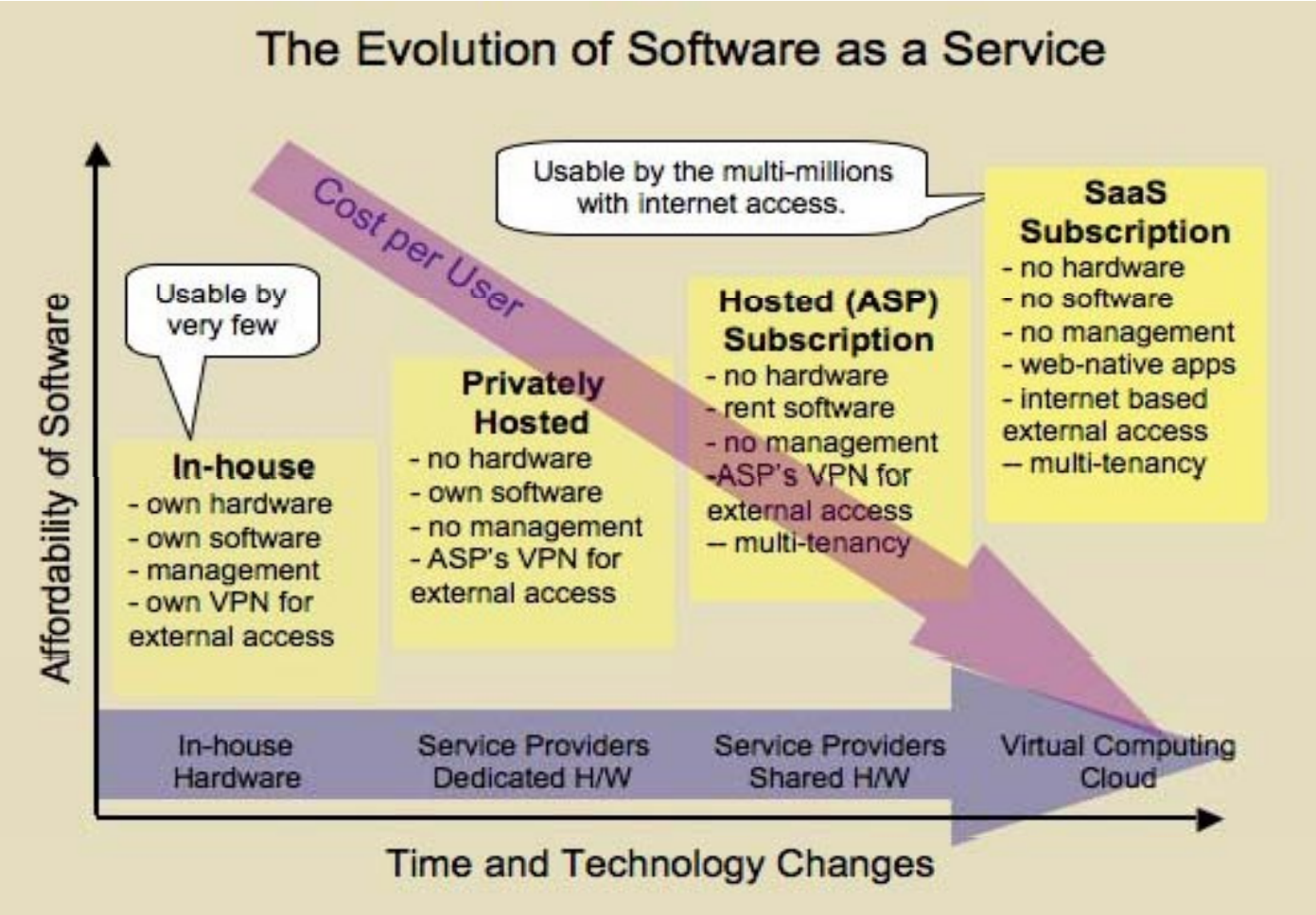

Software as a Service (SaaS)

- 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

  - Requirements: customizability and user data hosting

SaaS

- SaaS evolution



Cloud Economics

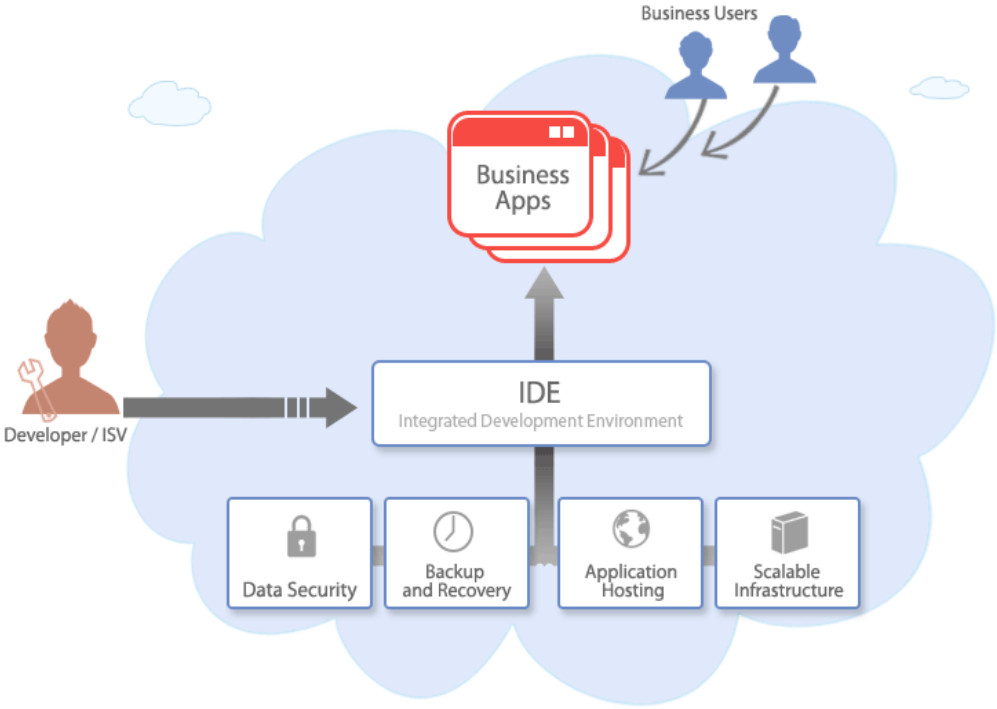
- Migrating all UOIT students to Gmail

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

Platform as a Service (PaaS)

- A supporting layer allowing users and developers to focus on their tasks

- Provide development, testing, deployment, hosting and maintenance solutions for cloud applications

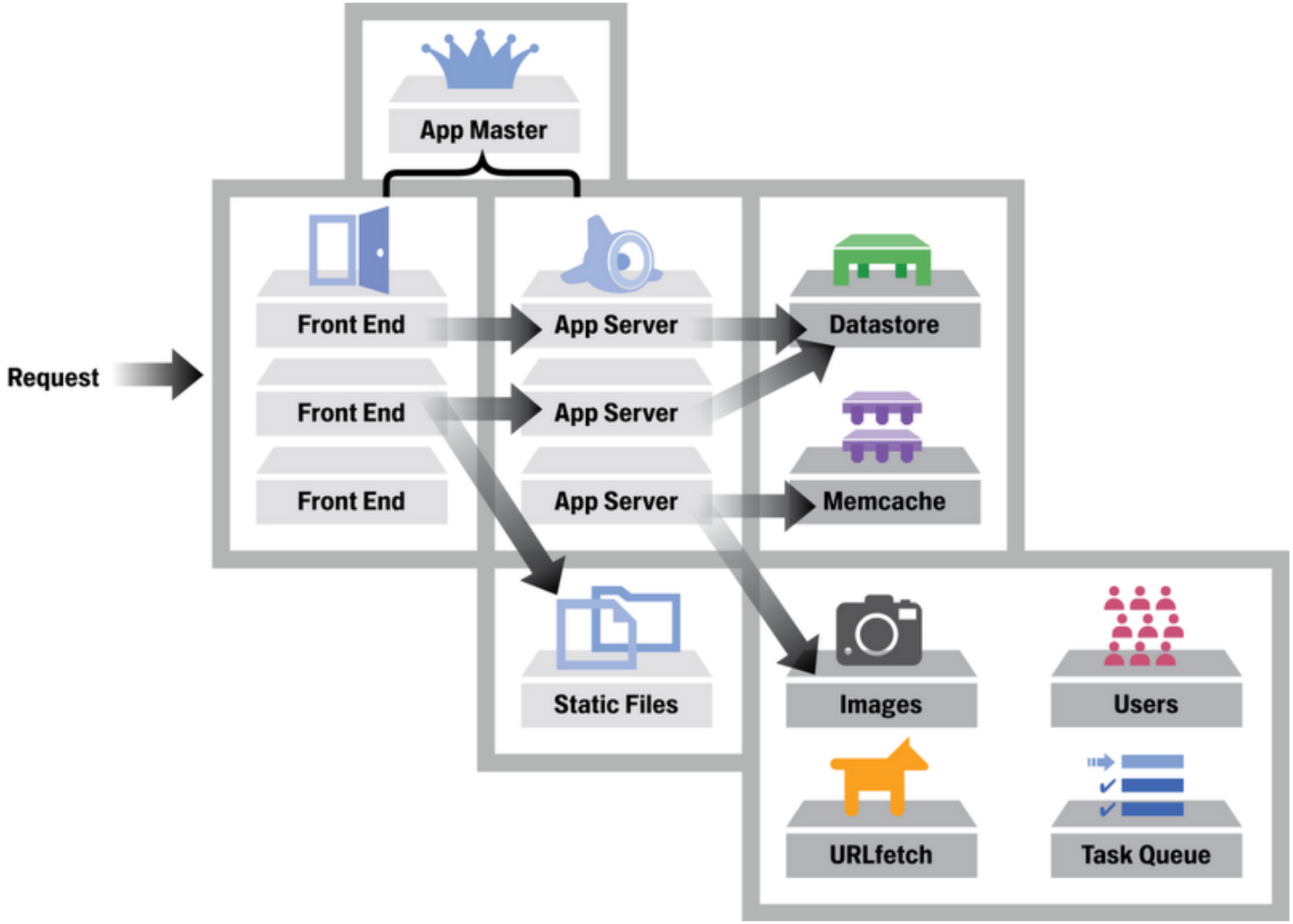


SaaS Provider: Heroku

- Heroku
  - Application platform

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

PaaS Provider: Google App Engine



Google App Engine

- 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)

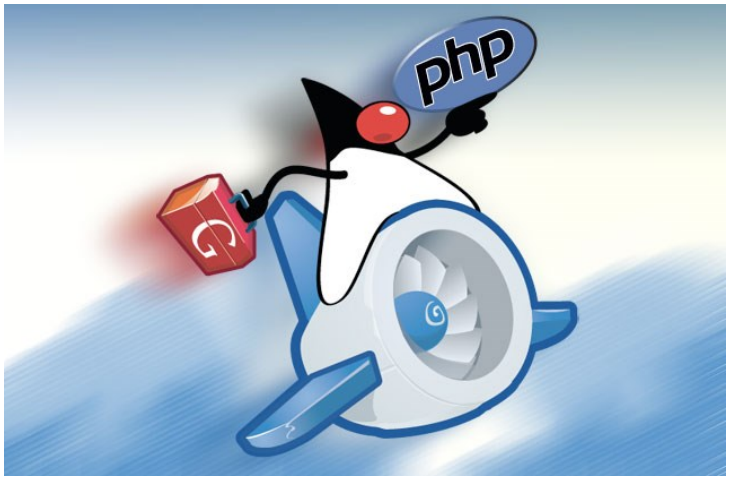
Google App Engine

- 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

- Developing a simple SaaS app:
  - Google App Engine
    - Java, Python, Go
  - Heroku (if time allows)
    - Supports many languages and frameworks



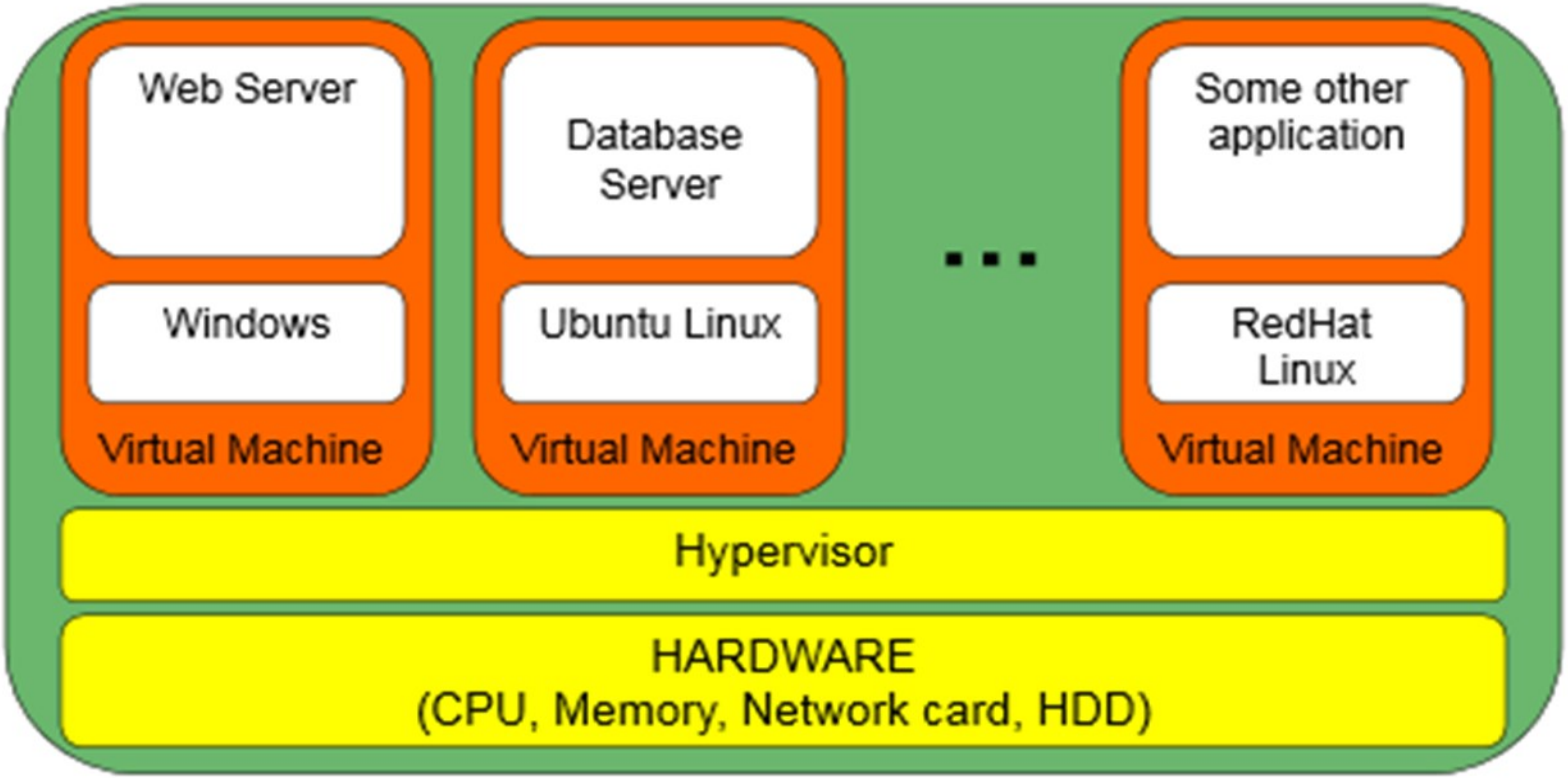


Characteristics of Cloud Computing

- 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)

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



IaaS

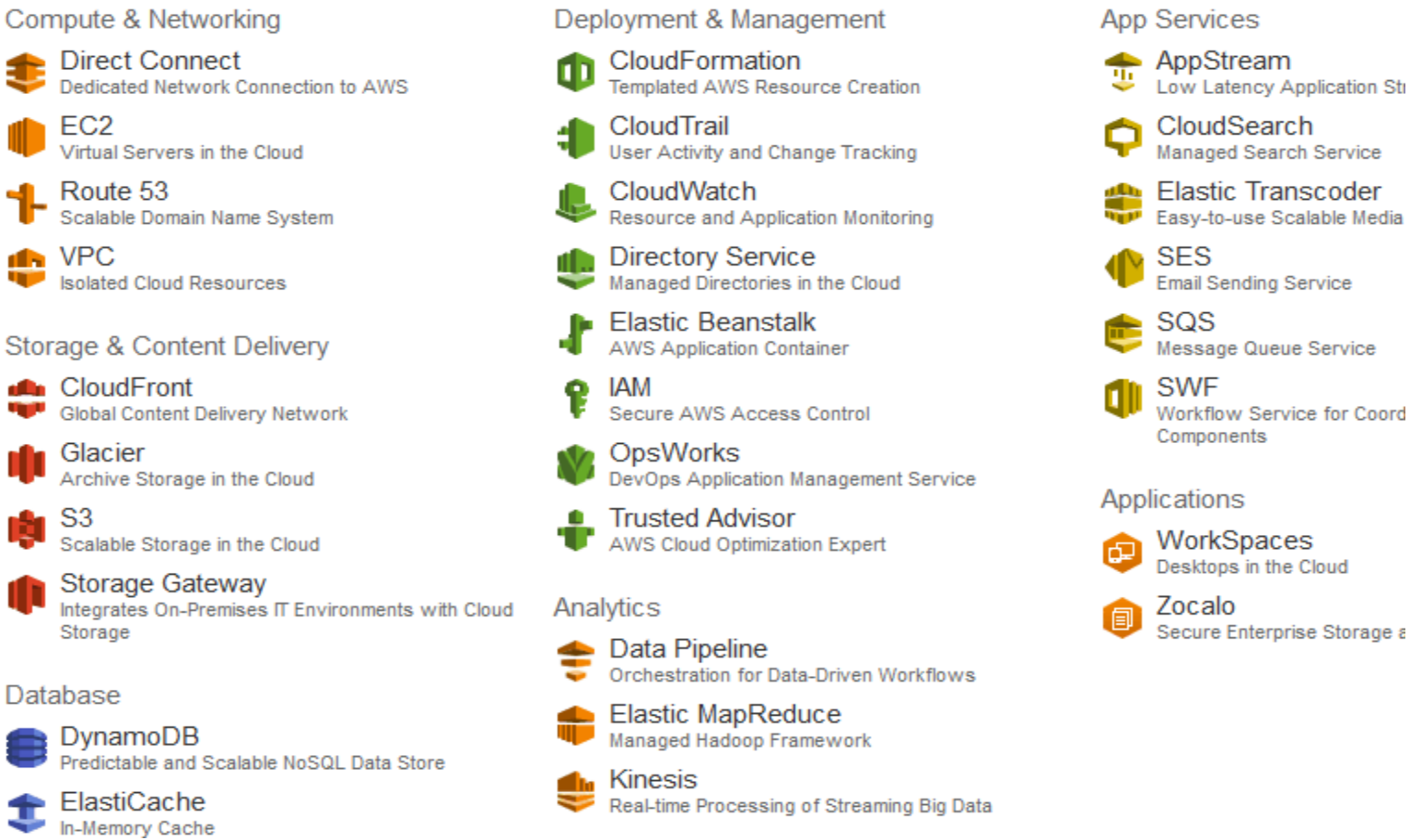
- IaaS focus on offering utility computing resources
- Requirements of IaaS
  - Scalability
    - System performance should remain the same (or at least similar) in small scale or large scale
  - Elasticity
    - Large amount of resource provisioning and deployment should be done in a short period of 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)

IaaS

- 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)

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



AWS

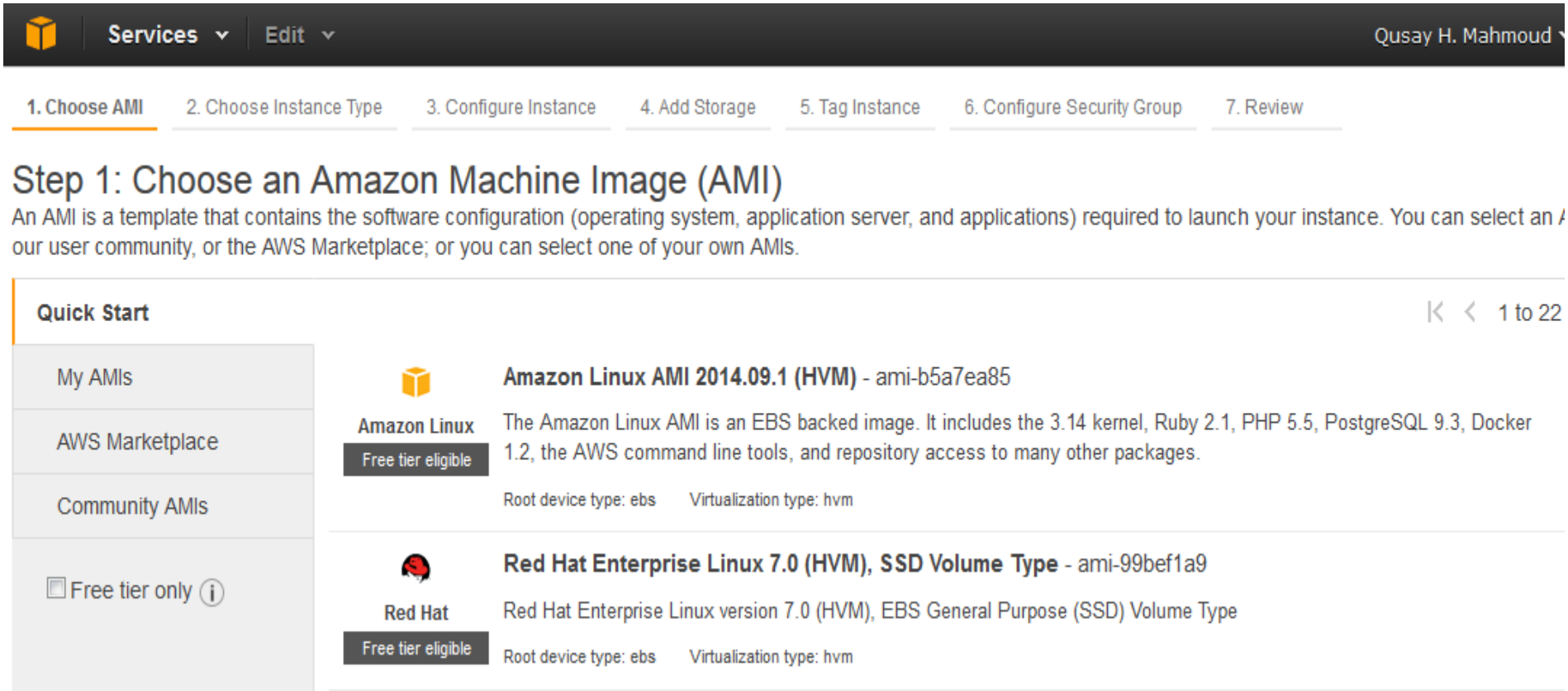
- Services that can be managed (created, monitored, terminated), over the Web
- Compute services
  - Elastic Compute Cloud (EC2): on-demand virtual machines (instances)
  - Elastic MapReduce (EMR): automatically starts Hadoop for parallel applications:
  - Auto scaling: seamless +/- number of EC2 instances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

AWS

- 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

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



Case Study: NY Times

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.

Magic EC2 / S3 Button

- 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

- 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

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

Case Study: CycleComputing

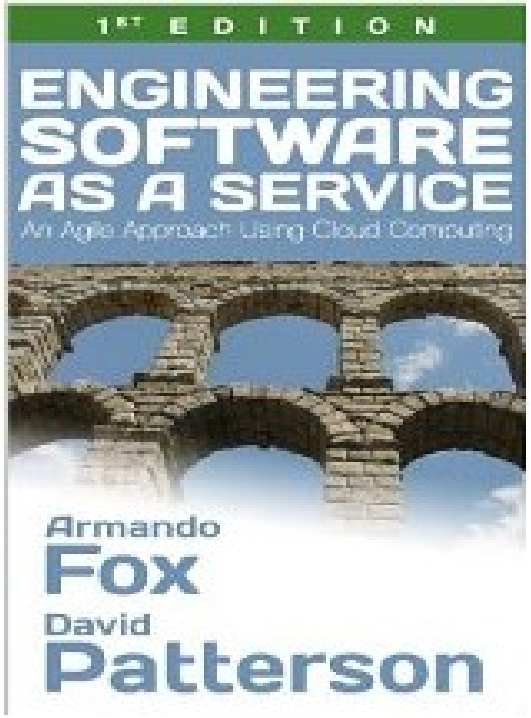
- Built a 50,000 core supercomputer called Naga using Amazon cloud infrastructure:
  - 6,742 instances
  - 51,132 cores
  - 58.78TB memory
  - Ran a Chemistry computational intensive job at $4,828.85/hr
  - They estimate the job used over 20 million $ in infrastructure

Teaching Cloud Computing

- 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

- 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

- 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

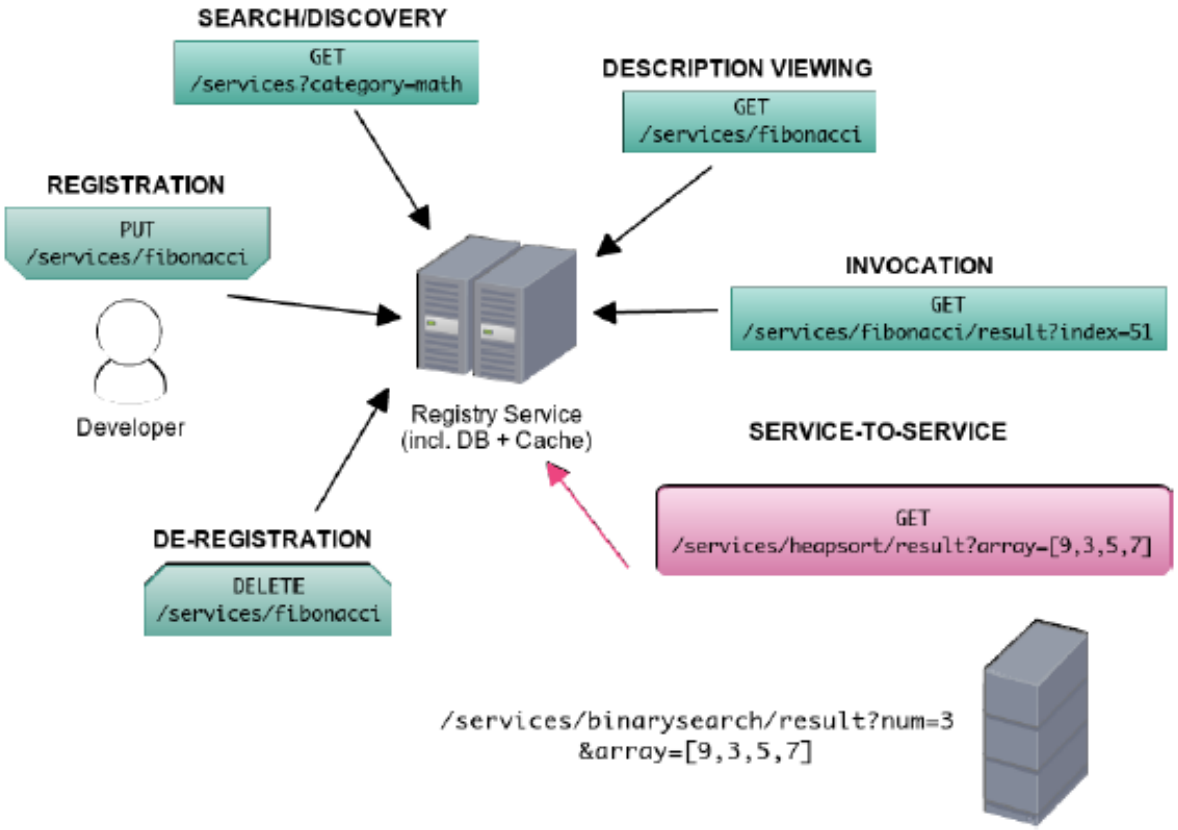
- Ok, I am convinced…where should I start?
  - Develop a new course on cloud computing (& big data)
- No space in the curriculum?
  - Integrate cloud computing (& big data) into existing courses:
    - Software Engineering
    - Security (hacking on VMs)
    - Databases
    - Networking
    - …others

Education and Research Grants

- 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…

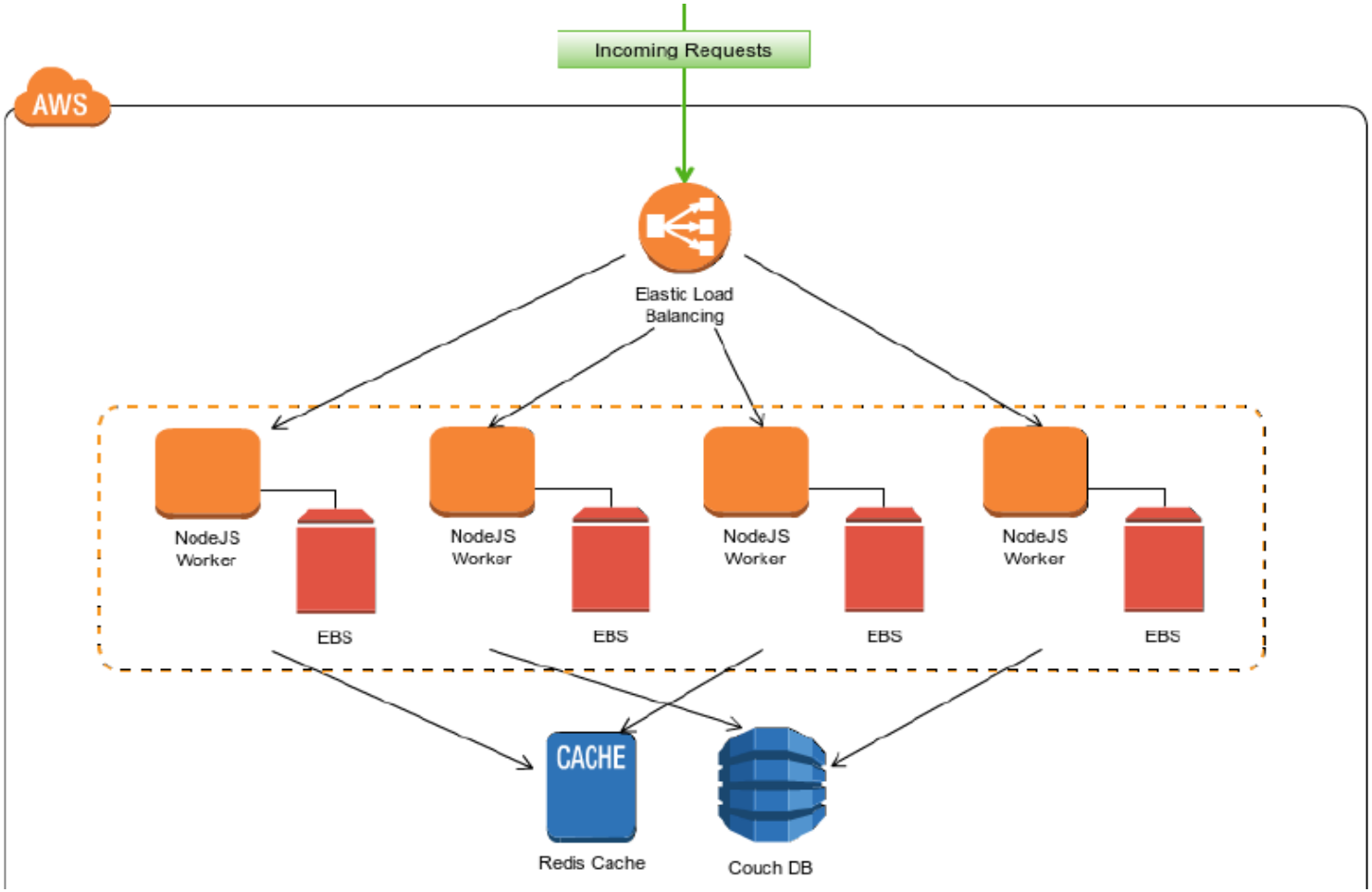
Cloud Computing for Research

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



Cloud Computing for Research

- Implementation of the platform in AWS

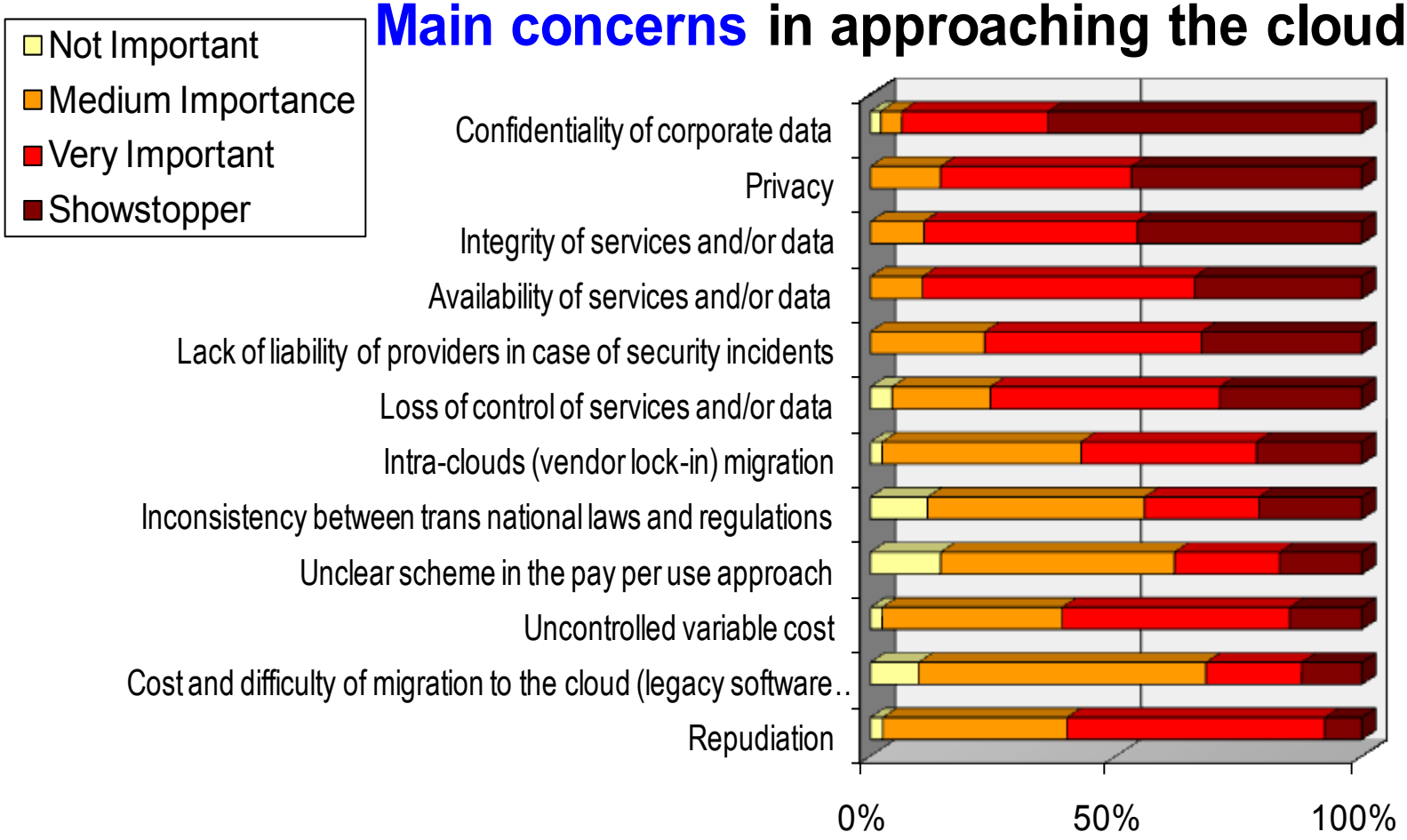


Research Opportunities

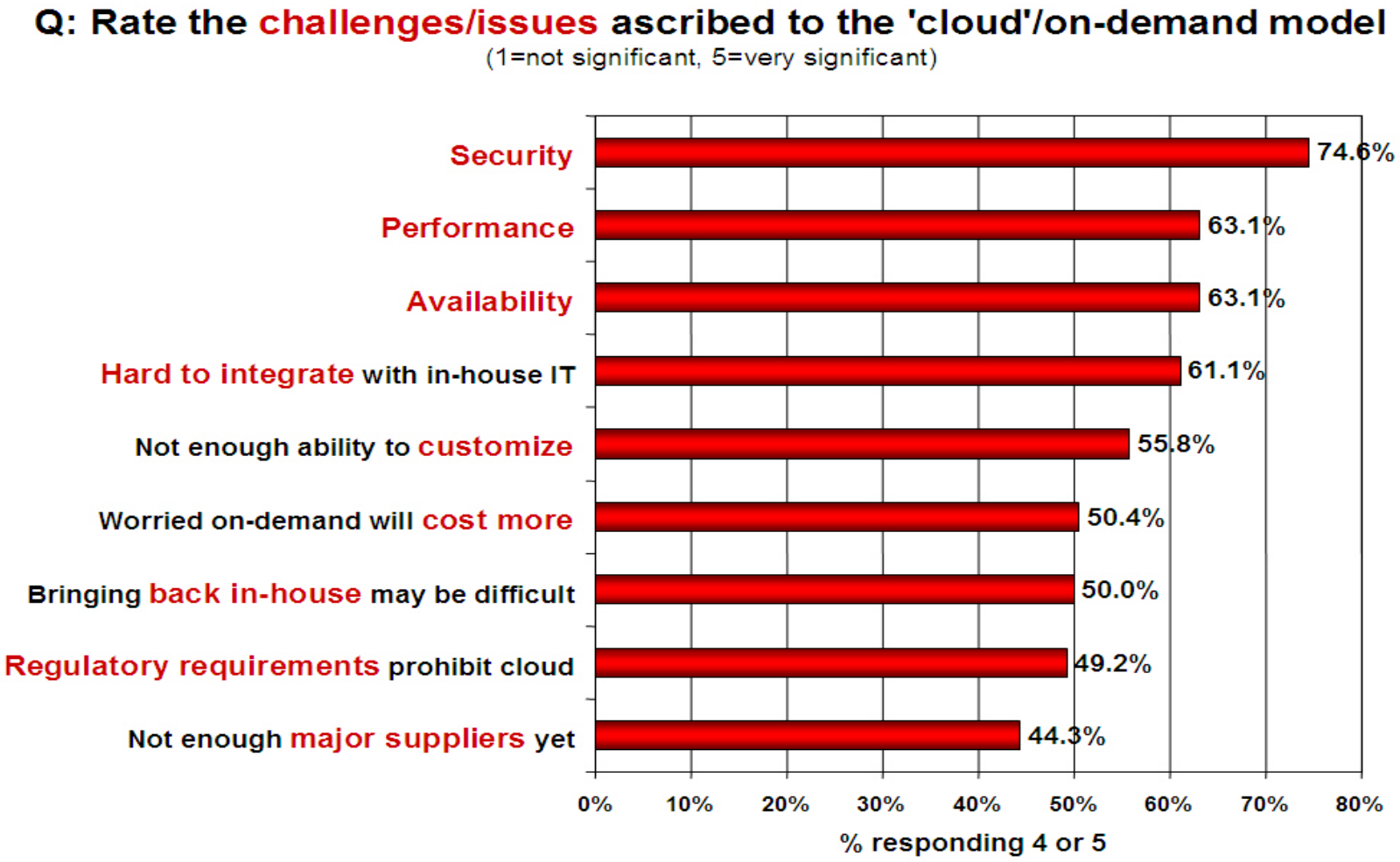
- Cloud computing (and SaaS) adoption in UAE
  - Or your institution
- Ankabut.ae
- EBTIC.org : Etisalat BT Innovation Centre
- Others…you may know about…
- Challenges:
  - Security, privacy, trust
  - Portability between providers (standards?)
  - Many others…

Main Concerns of Cloud Computing

- Results from 2009 survey by ENISA



Survey: Security is the major issue

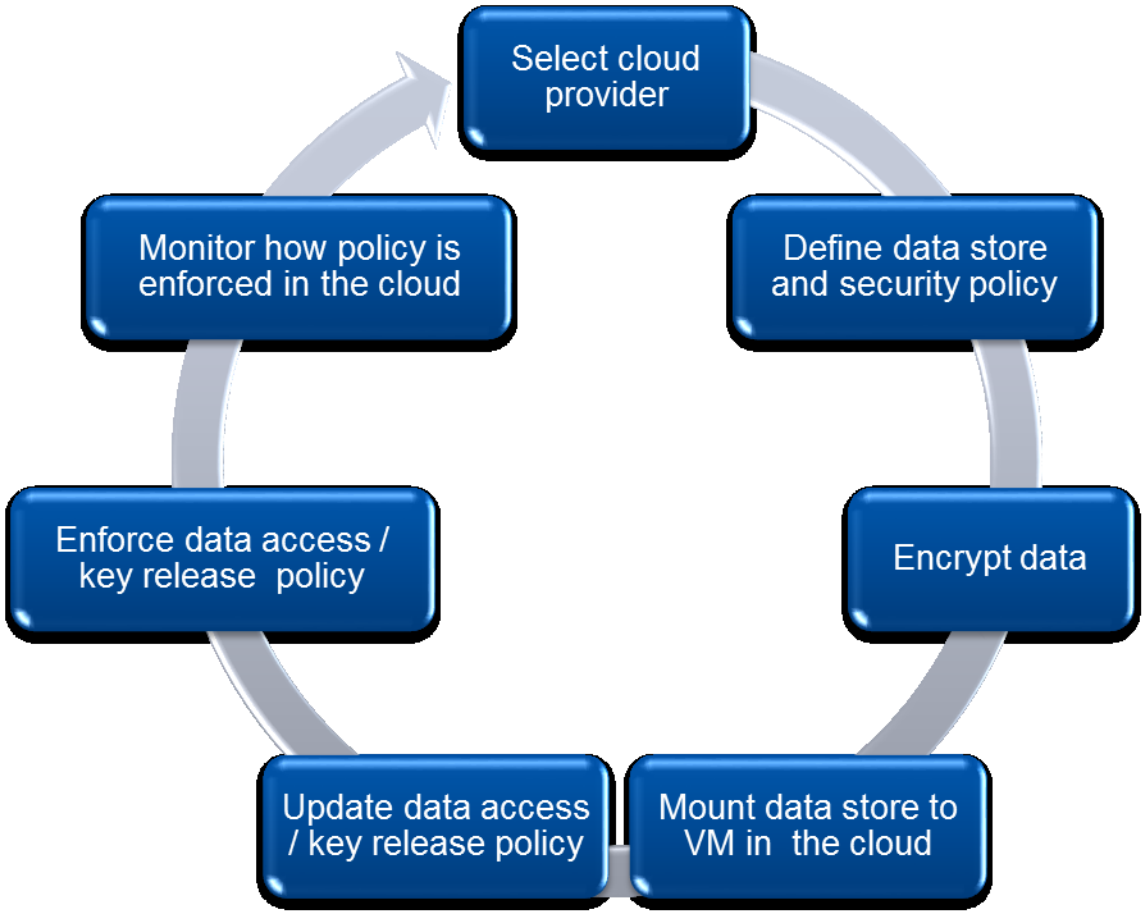


Source: IDC Enterprise Panel, August 2008 n=244

General Security Challenges

- 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



Big Data

Big Data

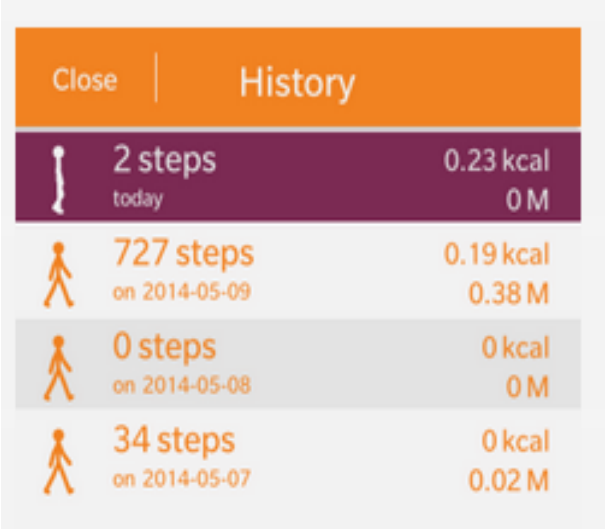
- Where does it come from?

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?

- 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



How much data does an airplane generate in one trip?

- Etihad Airways uses Big Data to reach its destination

  - 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

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

The Digital University

- 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

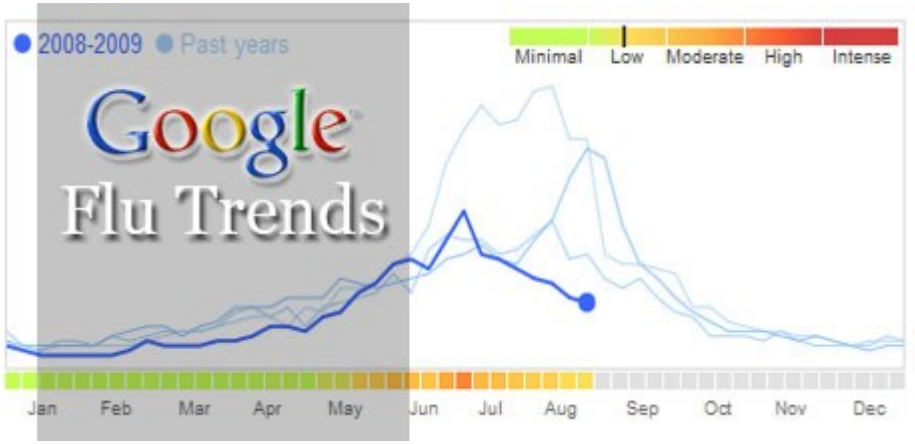
90% OF THE DIGITAL UNIVERSE IS
UNSTRUCTURED

IN 2011 THE DIGITAL UNIVERSE IS
300 QUADRILLION FILES

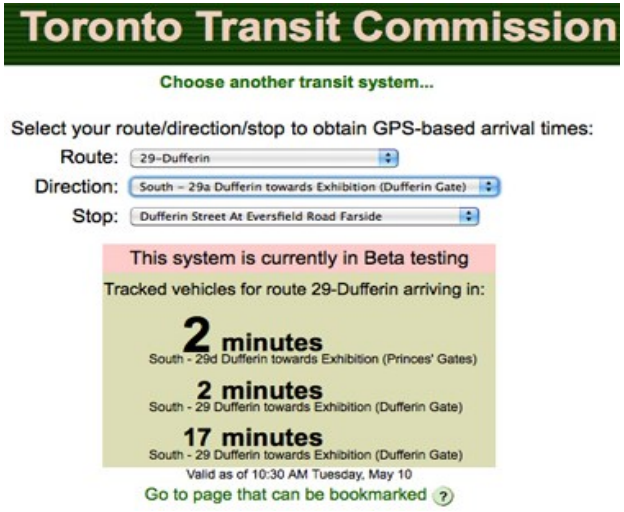
WIRED The New York Times Bloomberg Businessweek Forbes WALL STREET JOURNAL

Data-Driven Apps

- Google Flu / Correlate / Trends
- NextBus

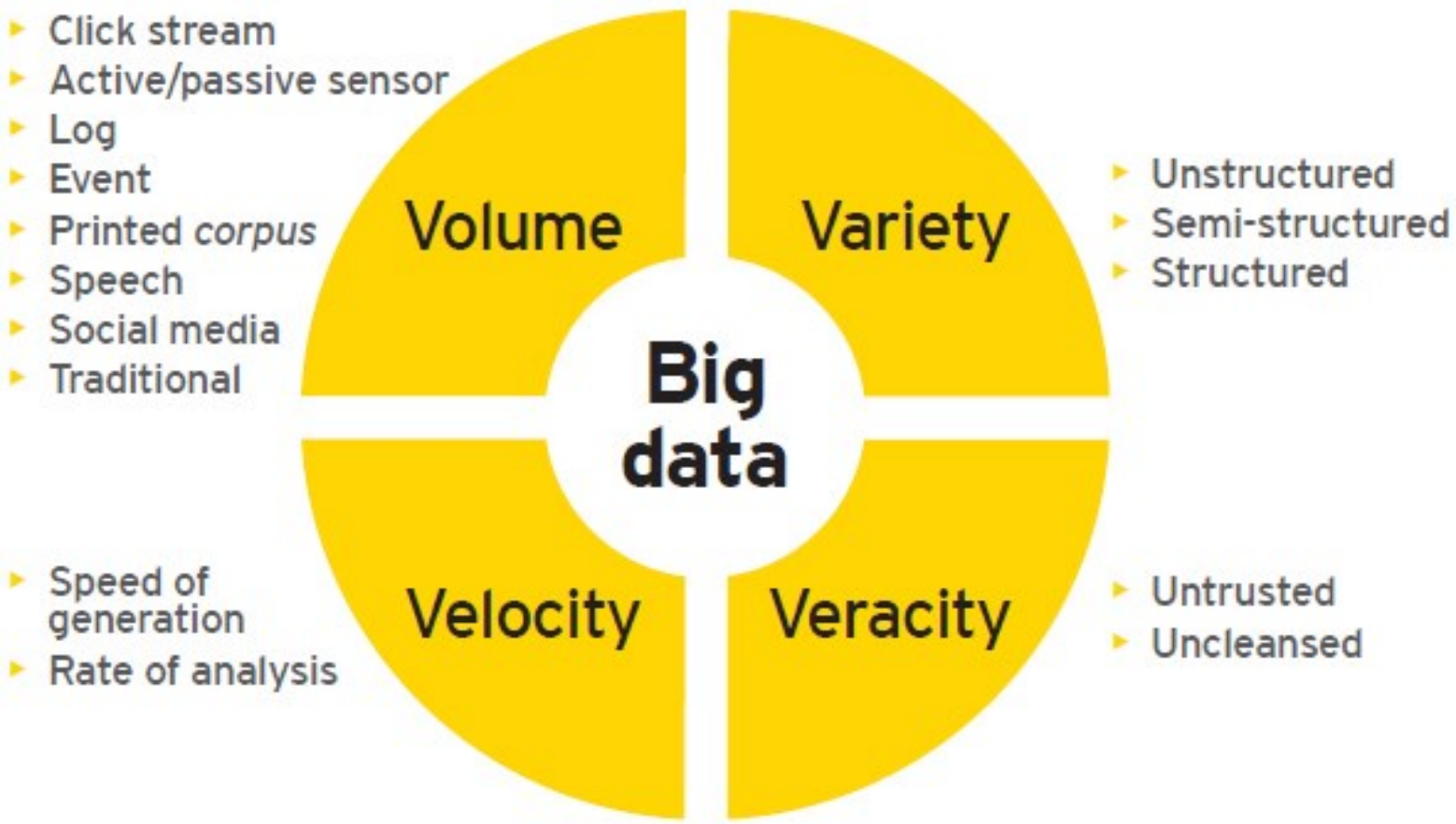


  - Data analytics examples:
  - Wefeelfine
  - Google correlate
  - Google flu



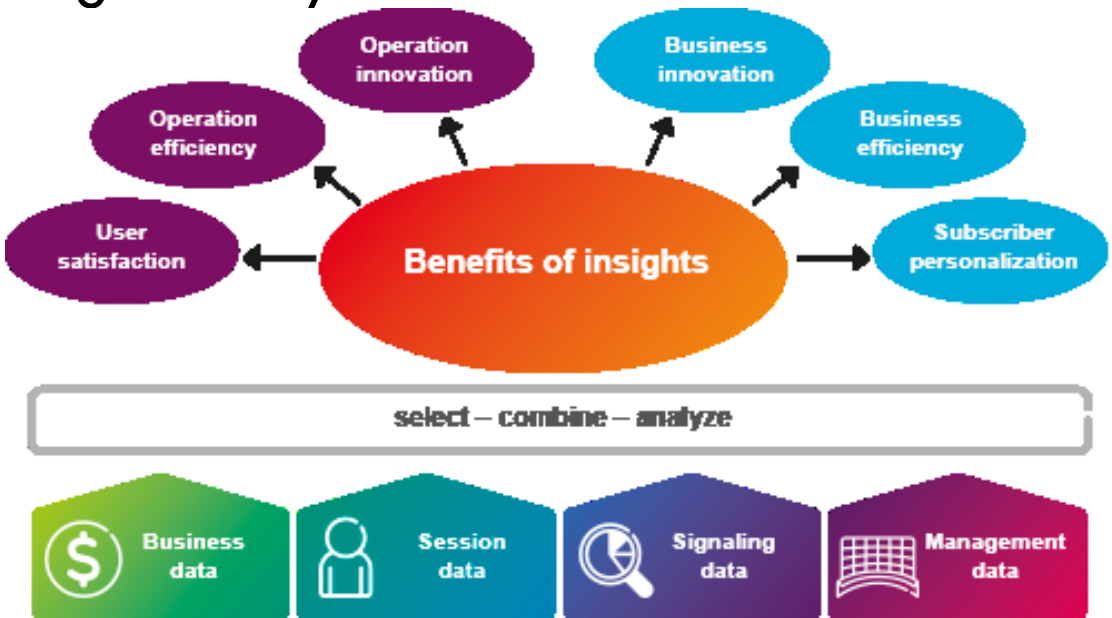
Big Data

- The four V's of big data



Big Data Analytics

- 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

- 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?

- A salable 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

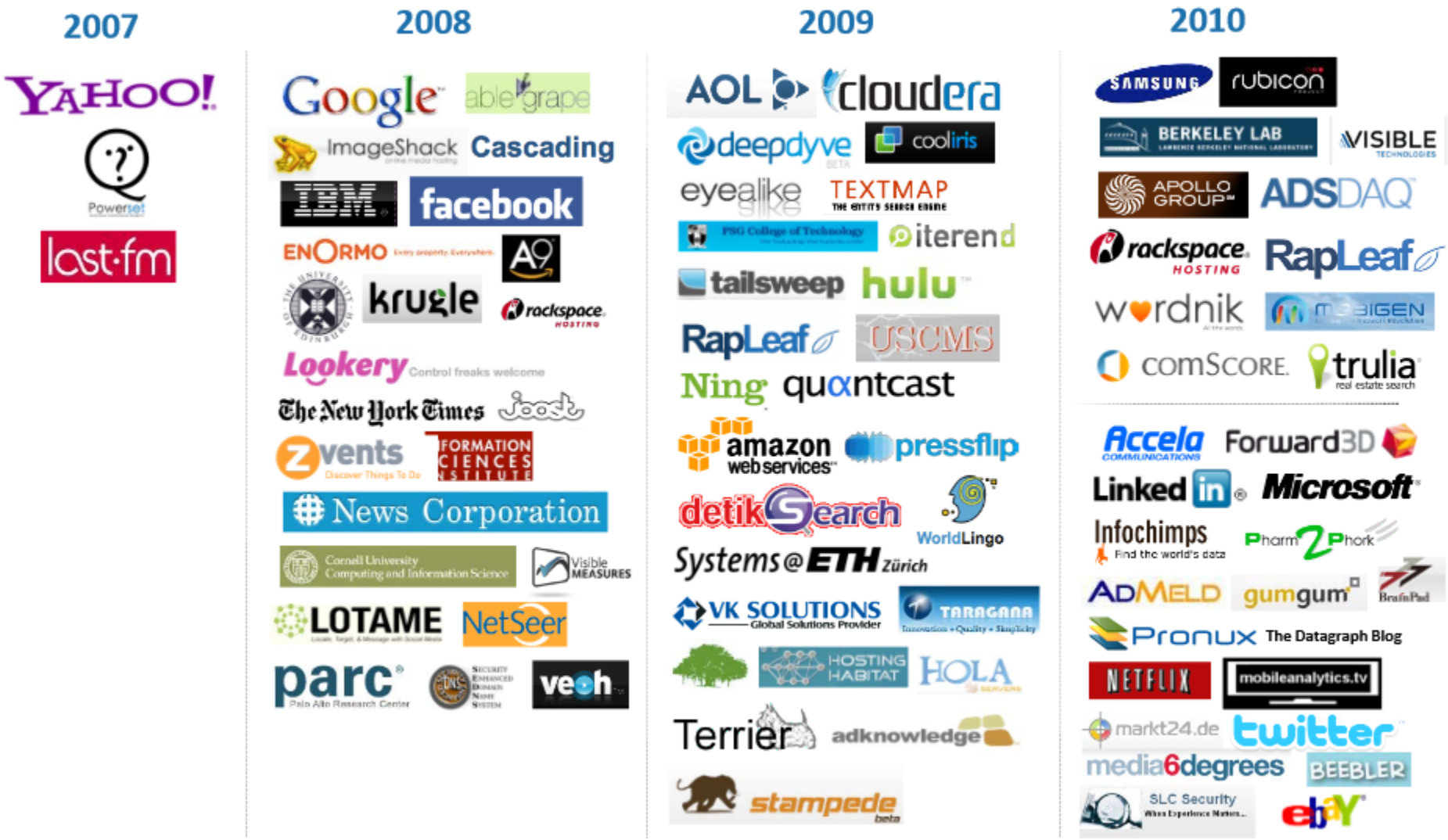
Why Hadoop?

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.

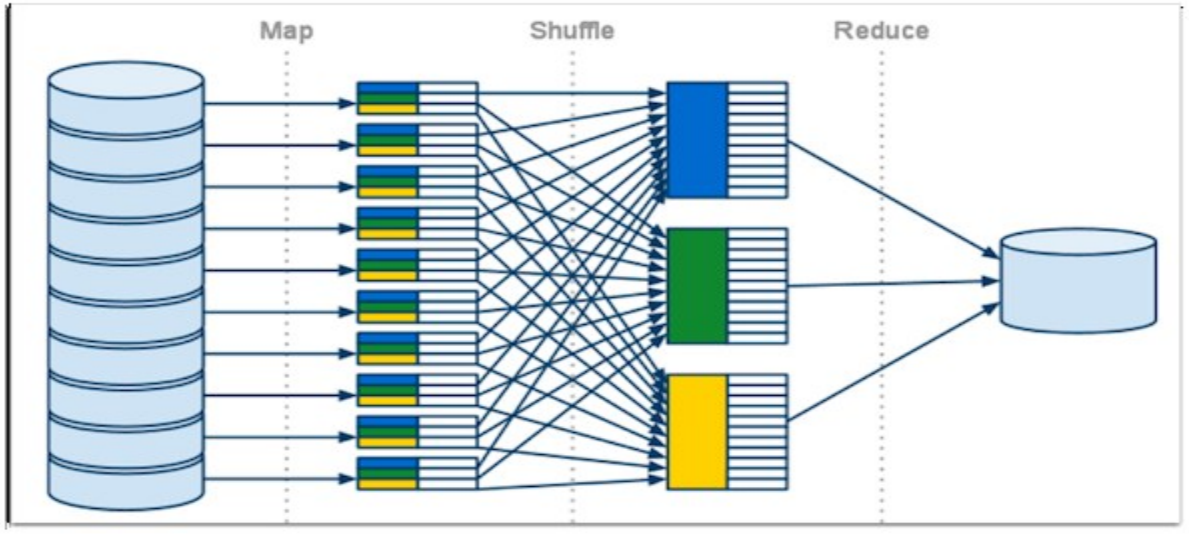
Gartner

Hadoop Adoption



Google MapReduce

- 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

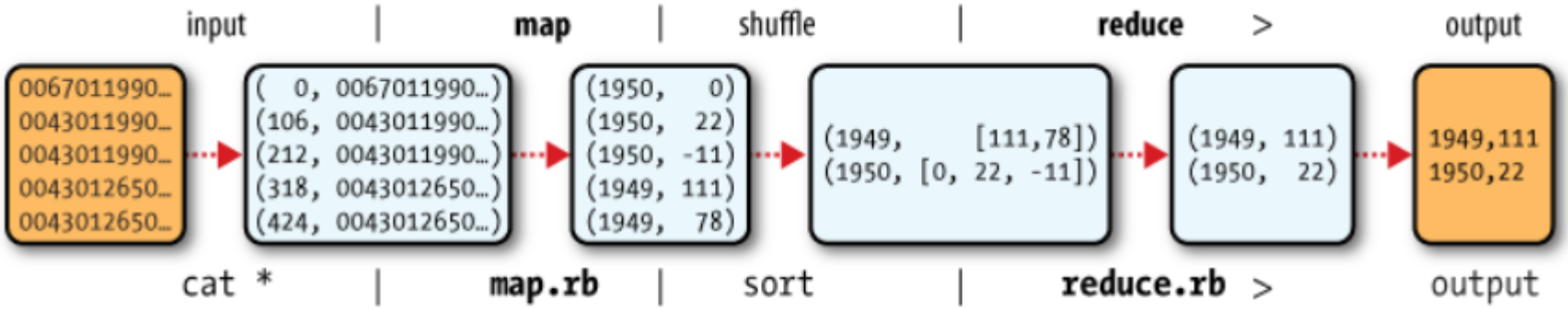


MapReduce Programming Model

- 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

- Man/min temperature for the last century?

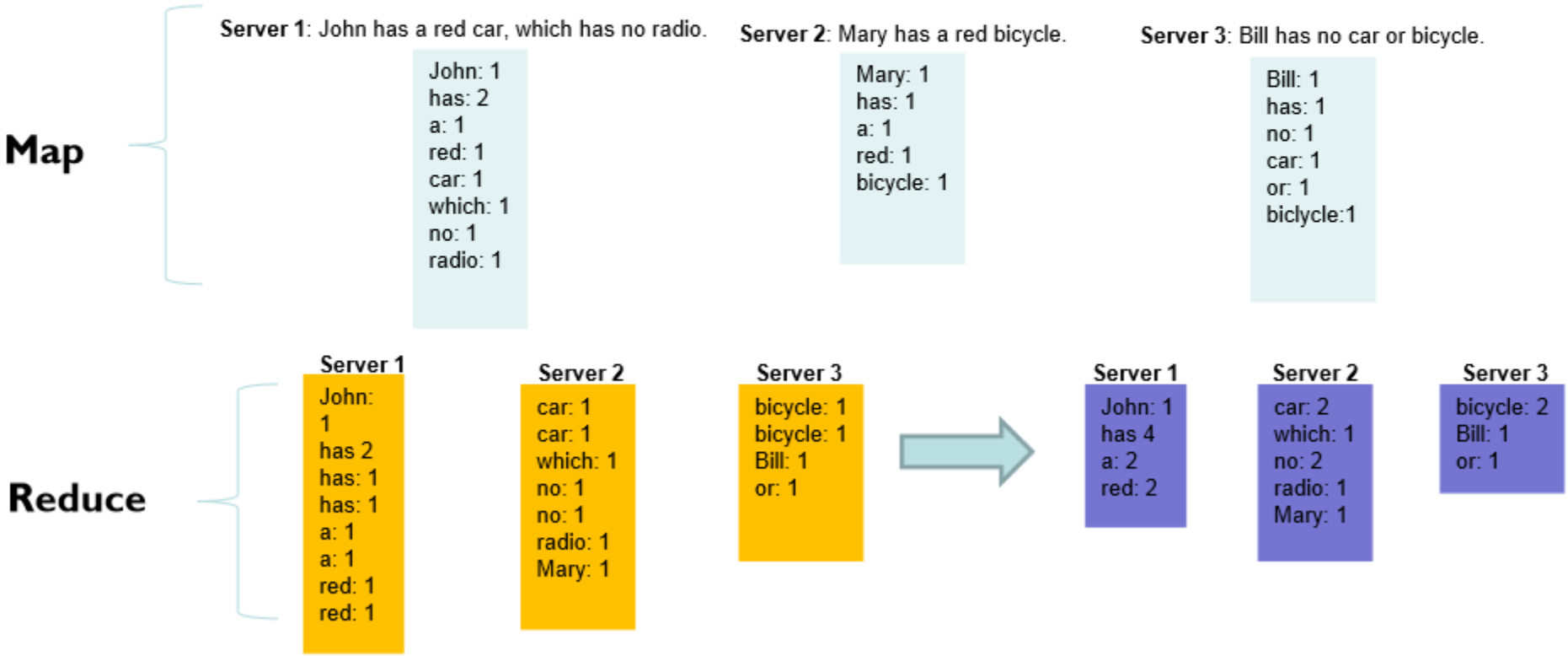


What was the max/min temperature for the last century?

Example e

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

- 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

- "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

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

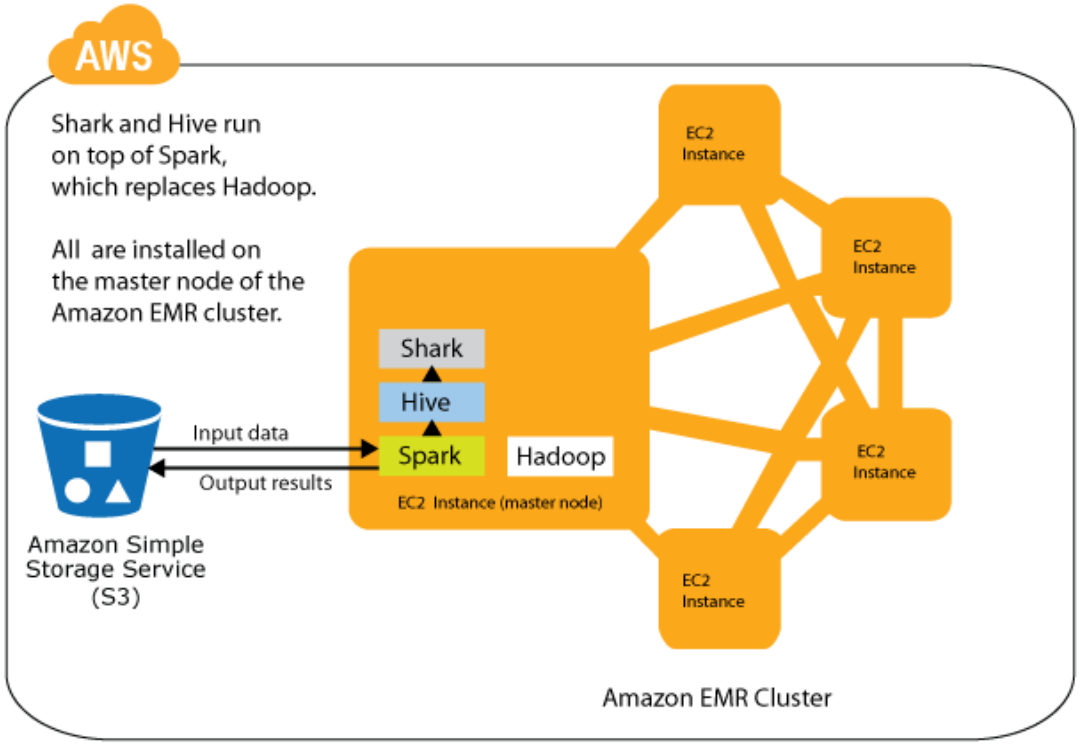
```
Reduce(String key,  values):
  // key: a word, same for input and output
 // values: a list of counts
int result = 0;
for each v in values:
    result = result + value;
Emit(result);
```

Example 3

- 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

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

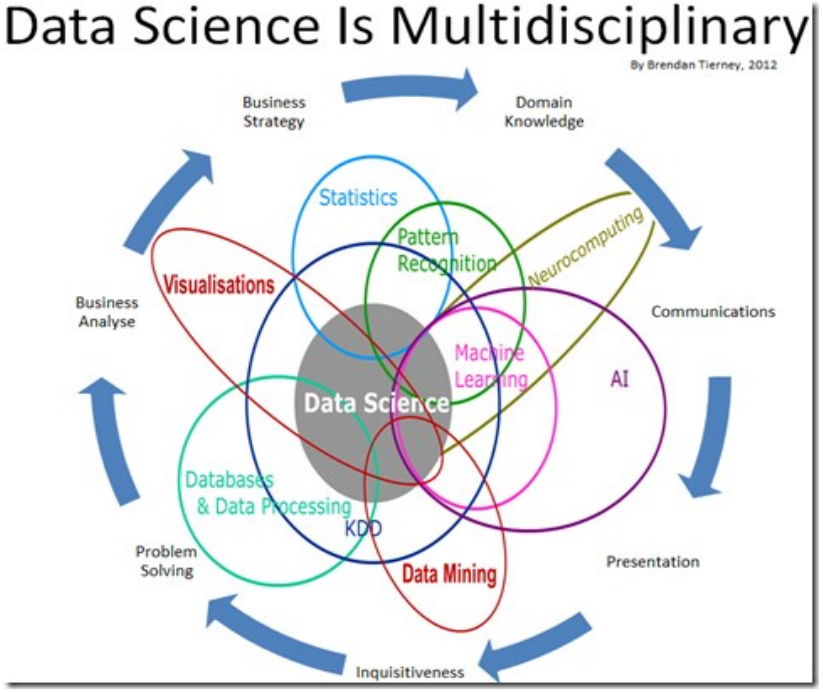


Big Data Security & Privacy Issues

- Discussion

Teaching

- 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?

Research

- 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

- 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

- 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