Week 11 - Wednesday

COMP 4290

Last time

- What did we talk about last time?
- Finished database reliability and integrity
- Sensitive data
- Database inference

Questions?

Project 3

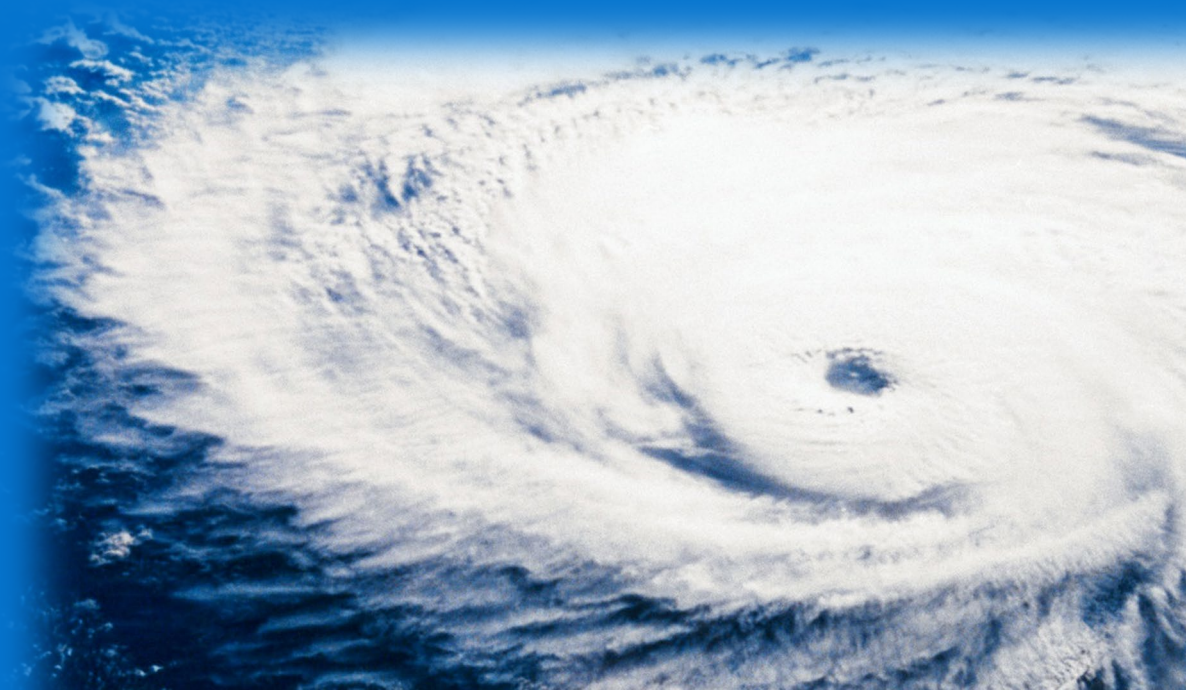
Assignment 4

Anu Regmi Presents

Data Mining

Data mining

- What do Walmart, hurricanes, and Pop-Tarts have to do with one another?
- A 2004 NY Times article says that Walmart's analysis shows the demand for strawberry Pop-Tarts goes up by a factor of 7 before a hurricane makes landfall
- But the top selling item is beer



Walmart

Save money. Live better.

Data mining

- **Data mining** means looking for patterns in massive amounts of data
- These days, governments and companies collect huge amounts of data
- No human being could sift through it all
- We have to write computer programs to analyze it
- It is sort of a buzzword, and people argue about whether some of these activities should simply be called **data analysis** or **analytics**

What is data mining?

- We have huge databases (terabytes or petabytes)
- Who is going to look through all that?
  - Machines of course
- **Data mining** is a broad term covering all kinds of statistical, machine learning, and pattern matching techniques
- Relationships discovered by data mining are probabalistic
  - No cause-effect relationship is implied

What can you do with it?

- It is a form of machine learning or artificial intelligence
- At the most general, you can:
  - **Cluster analysis:** Find a group of records that are probably related
    - Like using cell phone records to find a group of drug dealers
  - **Anomaly detection:** Find an unusual record
    - Maybe someone who fits the profile of a serial killer
  - **Association rule mining:** Find dependencies
    - If people buy gin, they are also likely to buy tonic

Privacy

- Social media providers have access to lots of data
- Facebook alone has details about more than 3 billion people
- Can they find hidden patterns about your life?
- Should they inform the police if they think they can reliably predict crime?
- What about data the government has?
- For research purposes, some sets of "anonymized" data are made public
  - But researchers often discover that the people involved can be discovered anyway

Data mining issues

- Privacy issues are complex
  - Sharing data can allow relationships to become evident
  - These relationships might be sensitive
- Integrity
  - Because data mining can pull data from many sources, mistakes can propagate
  - Even if the results are fixed, there's no easy way to correct the source databases
- Data mining can have false positives and false negatives

Cloud Computing

Cloud computing

- **Cloud computing** are flexible, Internet-based services that gives users access to computational resources on demand
- Cloud computing allows small companies to store and process data without the up-front costs of a data center
- Cloud computer services are growing rapidly, fueled by:
  - High-speed networking
  - Low cost computers and storage
  - Hardware virtualization technology
  - Demand for AI

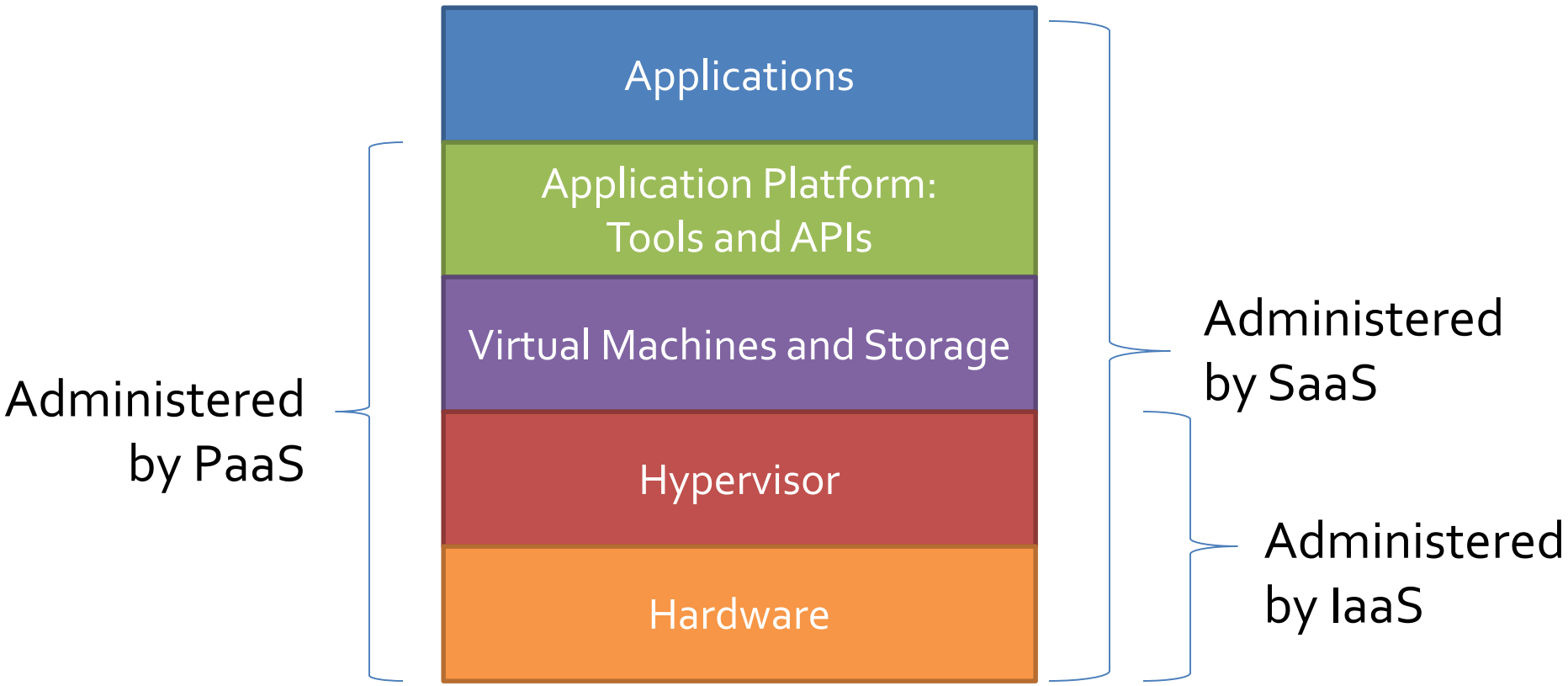
Defining characteristics

- Since cloud computing is a buzzword, we want to define clouds as having five characteristics:
  1. **On-demand self-service:** You can ask for more resources
  2. **Broad network access:** You can access services from lots of platforms
  3. **Resource pooling:** The provider has lots of stuff for you to use that can be dynamically assigned
  4. **Rapid elasticity:** Services can quickly and automatically be scaled up or down
  5. **Measured service:** You pay for computing like a utility

Service models

- **Infrastructure as a Service (IaaS)**
  - Processing, storage, and networks are in the cloud
  - You get (virtual) machines, but you're responsible for what's on them
- **Platform as a Service (PaaS)**
  - Languages, tools, and APIs are provided
  - You have to develop applications
- **Software as a Service (Saas)**
  - You get everything
  - You're using software and doing computations, but it's happening in the cloud

Another look at service models



Deployment models

- **Private cloud:** the cloud infrastructure is operated by and for the owning organization
- **Community cloud:** the cloud is shared by organizations usually with a common goal
- **Public cloud:** owned and operated by for-profit companies that make the service available to everyone
- **Hybrid cloud:** two or more clouds connected together

Moving to the cloud

- Should your business move to the cloud?
- There are steps you should take to determine the risk and value of doing so:
  - Identify assets you want to move to the cloud
  - Determine what additional vulnerabilities you will have on the cloud
  - Estimate the likelihood that those vulnerabilities will be exploited
  - Compute expected loss
  - Select new controls
  - Project total savings
- It may or may not save you money to move the cloud

Picking a provider

- Which model should you use?
- Even if you want public, there are many choices:
  - Amazon Web Services
  - Google App Engine (PaaS)
  - Google Compute Engine (IaaS)
  - Microsoft Azure (PaaS and IaaS)
- Important issues:
  - Authentication and access control
  - Encryption
  - Logging
  - Incident response
  - Reliability
- **Vendor lock-in** makes it hard to change providers

Cloud security tools

- Just using the cloud **can** have security benefits
  - Geographic diversity
  - Platform diversity
  - Infrastructure diversity
- Cloud platforms often support **mutual authentication**
- Cloud storage
  - There are risks when you share data on a platform
  - Consider how sensitive the data is
  - Consider how data sharing will be done
  - Are there laws or other regulations that apply?
  - Side channel attacks may be possible against other users of the same cloud

Huge Dropbox mistake

- Dropbox is a popular cloud service for backing up and synchronizing data
- On June 19, 2011, a bug in their software accepted *any* password for *any* account
- Dropbox said that files would be encrypted using the user password
  - But they weren't!
- When using a cloud service, it pays to look into the details

Cloud identity management

- Managing identities and authentication in a cloud can be challenging:
  - There are many computers communicating with each other
  - A hybrid cloud may have different authentication requirements within it
- **Federated identity management** is sharing identity information across different trust domains
  - There are systems for it, but it's a complex problem
  - It can provide **single sign-on** capabilities

Securing IaaS

- IaaS gives the user a lot of control
  - In other words, ways to be unsecure
- IaaS hosts can usually be controlled in more ways than traditional hosts
  - Good because it allows for robust logging and monitoring
  - Bad because there are more vulnerabilities attackers can try
- If you delete a file, it might not be gone, and someone else might be using the same hardware
- Authenticate command line interfaces strongly
- Use virtual machines that will only run specific applications
  - **Application whitelisting**

Upcoming

Next time…

- Privacy laws
- Web privacy
- Ashley Gutierrez presents

Reminders

- Read Sections 9.1 – 9.5
- Work on Project 3
  - **Passwords and secret phrases due this Friday!**
- Start Assignment 4
- **Exam 2 is next Monday**
  - **We'll review on Friday**