ISA 414 – Managing Big Data

Lecture 19 – Cloud Computing and Storage

(*Part I*)

Dr. Arthur Carvalho

arthur.carvalho@miamioh.edu

MIAMI UNIVERSITY

Announcements

- Final course project
  - Description is available on Canvas
  - Groups of 4
    - I will form the groups later
    - Your preferences will be taken into account
  - Sample reports are available on canvas
    - Keep in mind that the requirements have changed over the years

Announcements

- Final course project
  - Phase 1: presentation of initial ideas
    - November 16th and November 18th
    - Each presentation should last for 15 (ISA 414)/20 (ISA 514) minutes, leaving about 5 minutes for questions
    - Each presentation should cover the following:
      - Business background
      - Problem definition
      - Proposed solution: a bird's-eye view of the proposed solution
      - Potential pros and cons (*e.g.*, anticipated problems)

Announcements

- Final course project
  - Phase 2: work on project + report
    - In-class group work on November 23rd and November 30th
    - There is no final presentation
    - Main deliverable: a report that identifies the relevant problem and describes the proposed solution
      - Deadline: December 5th
    - Several constraints and requirements
      - See the project description on Canvas

Lecture Objectives

- Review the midterm project

- Learn some basic concepts related to cloud computing and storage
  - Different cloud service models
    - IaaS – Infrastructure as a Service
    - PaaS – Platform as a Service
    - SaaS – Software as a Service

Lecture Instructions

- Download the notebook "*Lecture 19.ipynb*" available on Canvas
  - Open that file with VS Code

Big Data

- Managing big data
  - This course:
  1. Past: **techniques** to effectively collect and analyze potentially unstructured data sets
  2. Present: big-data enablers, *i.e.*, modern **technologies** to address the challenges brought by big data
     - Cloud computing and storage
     - Distributed file systems (HDFS)
     - Parallel, distributed computing paradigms (Spark)

Big Data

- The volume and velocity aspects of big data bring new, complex challenges
  - How to store the data?
  - How to quickly capture and analyze the data?
- Several companies are embracing the challenges
  - Recent years have seen the rise of several powerful technologies and paradigms
    - *E.g.,* cloud computing/storage and Hadoop

Cloud Technologies

- For our purposes:
  - Cloud = IT infrastructure and applications on rent over the internet (*i.e.*, "*someone else's computer*")
- Let's consider a few things first before talking about cloud service models
  - To a certain degree, IT infrastructure is a commodity
    - Setting up such an infrastructure in-house requires:
      - (Disposable) Hardware
        - Computing devices
        - Storage devices
        - Network
        - …

Cloud Technologies

- For our purposes:
  - Cloud = IT infrastructure and applications on rent over the internet (*i.e.*, "*someone else's computer*")
- Let's consider a few things first before talking about cloud service models
  - To a certain degree, IT infrastructure is a commodity
    - Setting up such an infrastructure in-house requires:
      - People (experts)
        - IT/management team
        - Security team
      - High up-front capital investment

Cloud Technologies

- Maintaining an in-house IT infrastructure is neither cheap nor easy
  - Disposable hardware
    - A new computer today will be a mediocre computer in 3 years and an old computer in 6 years
  - People move
    - Jobs in high demand = many new offers
    - Costly training process

Cloud Technologies

- Maintaining an in-house IT infrastructure is neither cheap nor easy
  - Many startups or companies growing too fast often struggle keeping up with a modern infrastructure
- Alternative: cloud computing and storage
  - On-demand infrastructure
    - Computation anytime anywhere – whenever there is a demand
  - Pay to use someone else's IT infrastructure
    - Rental service: rent what you want, and return upon usage

Cloud Technologies

- Analogy
  - Would you buy (or build) a truck every time you have to move a piece of furniture?
    - Do you have the resources and/or skills?
    - Cost/benefit tradeoff

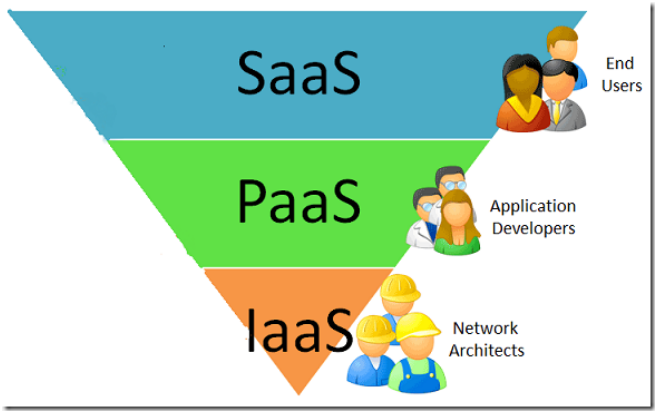
  - Would it make more sense to rent a truck, and return it when the moving is done?

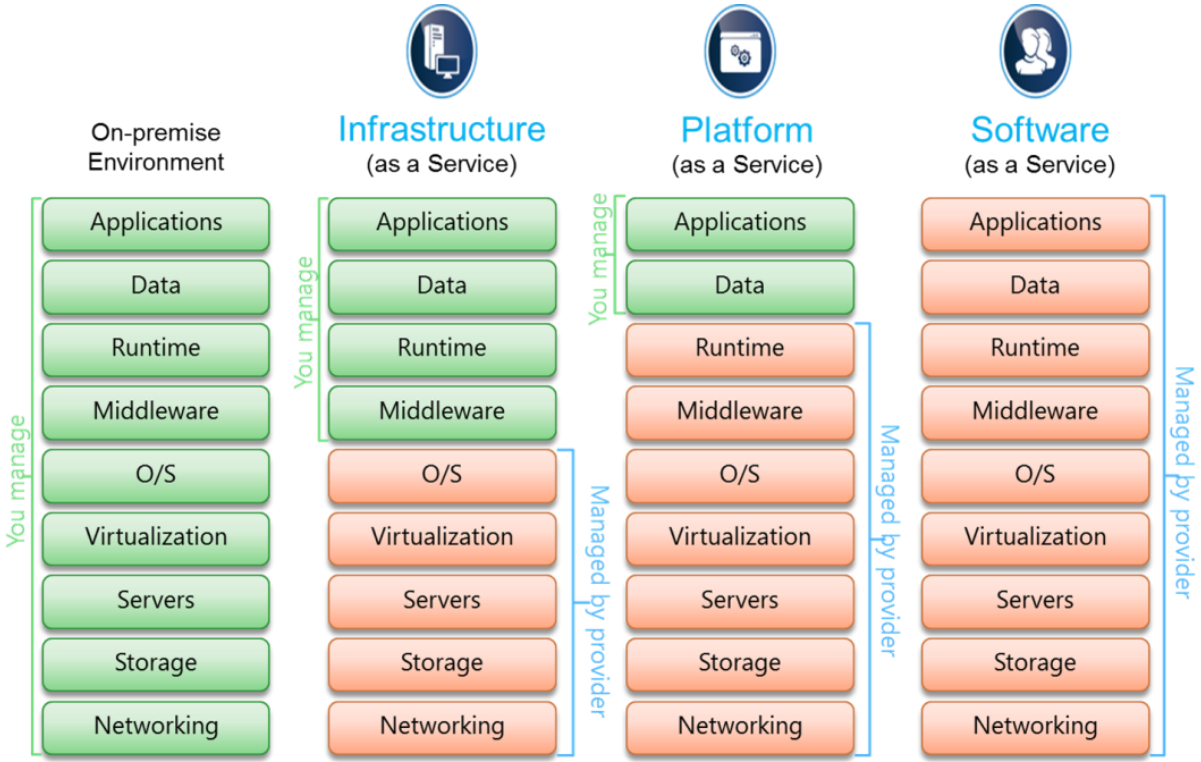
- Similarly, why building an IT infrastructure when you can rent one?

Cloud Technologies

- So far, we have focused our discussion on IT infrastructure
  - Infrastructure as a Service(IaaS)
  - Other cloud service models include renting:
    - Platform (Platform as a Service)
    - Application (Software as a Service)
  - Bottom line: there are different business models around the concept of cloud with different levels of engagement and servicing
    - Similar to rental agreements

Cloud Service Models

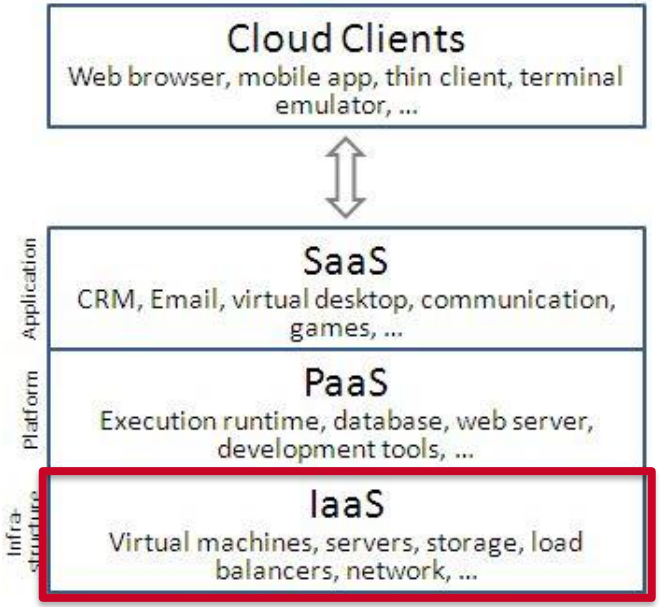




MIAMI UNIVERSITY

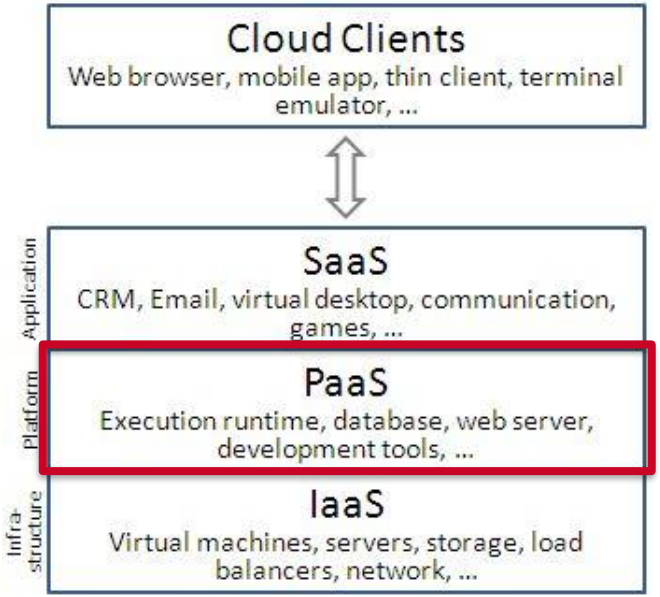
IaaS

- **IaaS**: Infrastructure as a Service
  - Service: hardware only
    - Bare minimum rental service
  - The client still has to install and maintain any required software or service



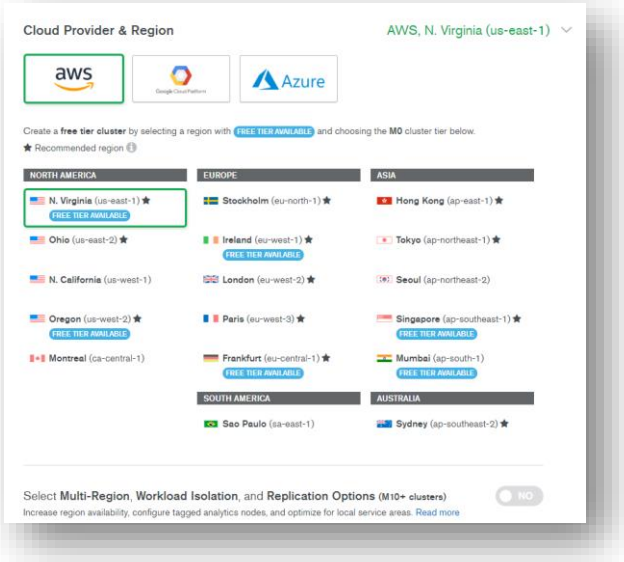
PaaS

- **PaaS**: Platform as a Service
  - Service: computing environment
    - Might include:
      - Programming languages and development tools
      - Web/database servers



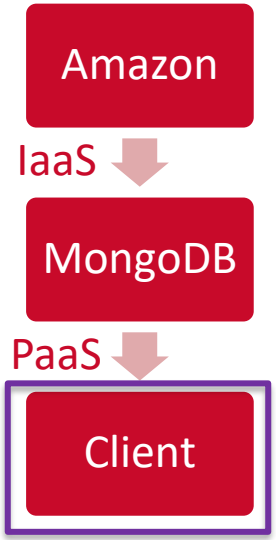
PaaS

- Example: cloud-based MongoDB
  - Recall Lecture 11
    - MongoDB database provided by MongoDB company



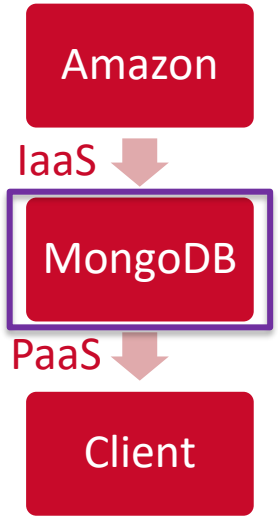
PaaS

- Example: cloud-based MongoDB
  - A client pays MongoDB for a database
    - PaaS (no need to think about OS, servers, *etc*.)
  - The client must still manage the database
    - *E.g.*, create users, set privacy constraints, *etc*.
  - The client can build apps/software that use the database (*e.g.*, Assignment 3)
    - No need to worry about the hardware infrastructure
    - Buy more storage space from MongoDB (company) as required



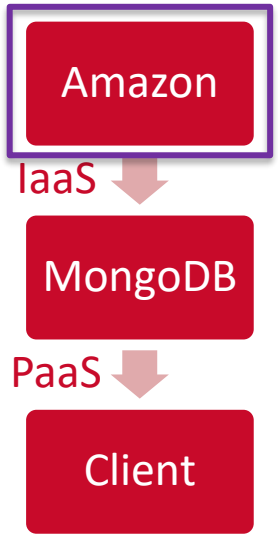
PaaS

- Example: cloud-based MongoDB
  - MongoDB (company) pays Amazon for using their computers, storage device, network
    - IaaS
  - MongoDB (company) manages the database infrastructure
    - Monitor storage capacity, database speed, security
    - Buy more storage and computational power from Amazon when needed (*e.g.*, too many new clients)



PaaS

- Example: cloud-based MongoDB
  - Amazon
    - Manages the hardware
      - Computers, storage devices, network, …
  - Business model: monthly payment
    - MongoDB (company) is the “man in the middle”
      - Easy way of making money



PaaS

- Example
  - Pricing scheme:
    (as of 07/14/2021)

Amazon Web Services (AWS) | Google Cloud Platform | Microsoft Azure

Available in most AWS regions with prices (in USD) varying across regions. Prices below are for the US East (N. Virginia) region.

DEDICATED CLUSTER
Standard

This line offers the most economical plans for production applications running on AWS.

Each Replica Set comes standard with 2 data-bearing nodes and high-availability via auto-failover.

	RAM	vCPUs	Storage	Monthly price
M1	2 GB	1	40 GB	$180
M2	4 GB	2	60 GB	$360
M3	8 GB	2	120 GB	$720
M4	16 GB	4	240 GB	$1,430
M5	32 GB	8	480 GB	$2,260
M6	64 GB	16	700 GB	$3,520
M7	122 GB	16	700 GB	$4,540

DEDICATED CLUSTER
High Storage

This line offers a higher storage-to-RAM ratio than our Standard line and is geared towards applications that need to store large amounts of data but have more modest performance requirements.

Each Replica Set comes standard with 2 data-bearing nodes and high-availability via auto-failover.

	RAM	vCPUs	Storage	Monthly price
M1	2 GB	1	75 GB	$220
M2	4 GB	2	150 GB	$440
M3	8 GB	2	300 GB	$880
M4	16 GB	4	600 GB	$1,760
M5	32 GB	8	1 TB	$2,730
M6	64 GB	16	1 TB	$3,790
M7	122 GB	16	1 TB	$4,810

DEDICATED CLUSTER
High Performance

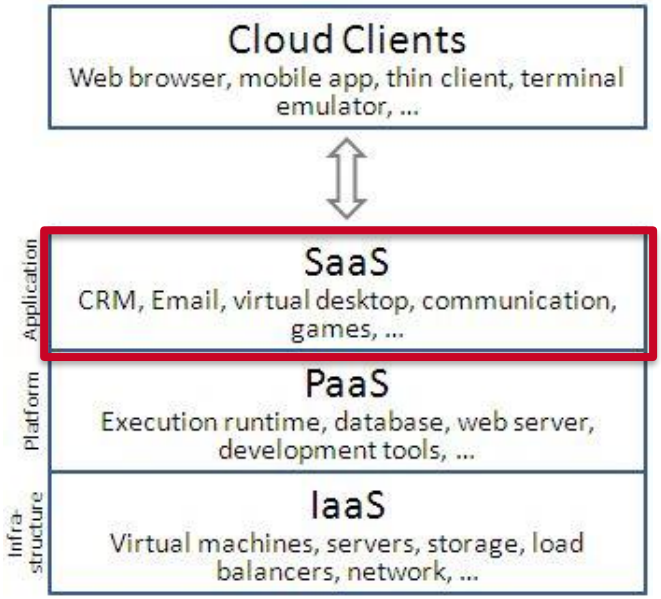
This line is designed for applications with the most demanding workloads, featuring local SSD storage with very high throughput and extremely low latency.

Each Replica Set comes standard with 2 data-bearing nodes and high-availability via auto-failover. A hidden 3rd node is included for increased durability and backups.

	RAM	vCPUs	Storage	Monthly price
M5	30 GB	4	850 GB	$3,750
M6	60 GB	8	1 TB	$5,890
M7	122 GB	16	1 TB	$10,130

SaaS

- **SaaS**: Software as a Service
  - Service: full software/service on demand
  - Cloud service provider takes care of the whole environment
    - Hardware + software
  - Example:
    - Dropbox, Google Docs, Microsoft Office online



XaaS

- XaaS: anything as a service
  - Marketing aaS
    - *E.g.*, IBM Watson Marketing Insights
  - Analytics aaS
    - *E.g.,* IBM Watson Analytics
  - …

Case Study

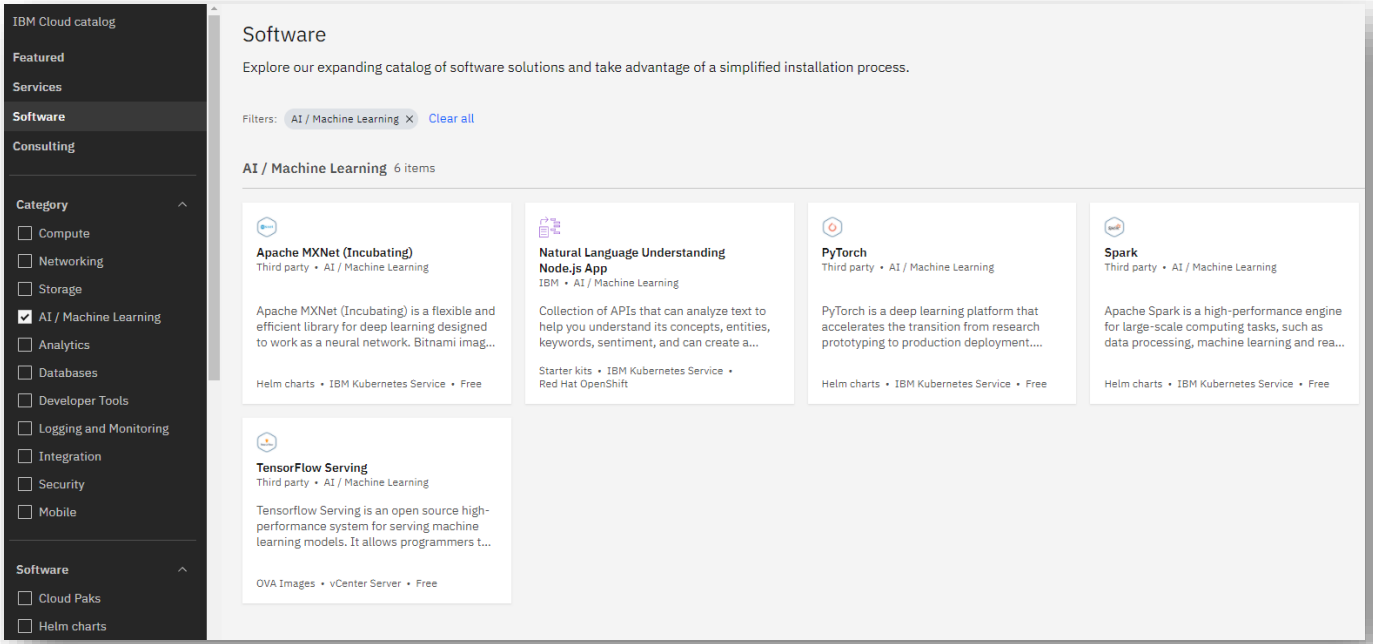
Sentiment Analysis with Natural Language Understanding: A SaaS Provided by IBM Cloud

IBM Cloud

- There are many cloud platforms that offer different SaaS, PaaS, IaaS
  - Microsoft Azure
  - IBM Cloud (formerly, IBM Bluemix)
  - Google Cloud
  - Amazon Web Services (AWS)
- We will be focusing on IBM Cloud
  - Until a few weeks before the beginning of the course, IBM Cloud was the only provider that did not require credit card info when creating free accounts
  - The above has changed ☹

IBM Cloud

- All the services provided by IBM Cloud
  - We will explore them in our next class



MIAMI UNIVERSITY

Natural Language Understanding

- Let's experience one IBM Cloud's service
  - Natural Language Understanding
    - Sentiment analysis
  - Go to https://www.ibm.com/demos/live/natural-language-understanding/self-service for a demo
    - We will soon request this service using Python
      - REST + JSON

Natural Language Understanding

- Business model:
  - Payment per call (+ tailored model)

(as of 07/14/2021)

Natural Language Understanding

Prices displayed for region: Dallas

Lite

30,000 NLU Items Per Month **Free**

1 Custom Model

Fixed API Rate Limit. See Standard plan for higher API Rate Limit

NOTE: A NLU item is based on the number of data units enriched and the number of enrichment features applied. A data unit is 10,000 characters or less. For example: extracting Entities and Sentiment from 15,000 characters of text is (2 Data Units * 2 Enrichment Features) = 4 NLU Items. A custom model refers to an annotation model developed with Watson Knowledge Studio.

The Lite plan gets you started with 30,000 NLU Items per month at no cost. This plan also enables use of one custom model published through Watson Knowledge Studio.

Standard

Unlimited NLU Items Per Month

You will be charged per NLU Item & per Custom Model

Increased API Rate Limit can be configured upon request

NOTE: A NLU item is based on the number of data units enriched and the number of enrichment features applied. A data unit is 10,000 characters or less. For example: extracting Entities and Sentiment from 15,000 characters of text is (2 Data Units * 2 Enrichment Features) = 4 NLU Items. A custom model refers to an annotation model developed with Watson Knowledge Studio.

This plan allows unlimited NLU Items and adds the ability to use multiple custom models published through Watson Knowledge Studio.

TIERS	PRICING
1 - 1	$800.00 USD/Custom Model Instance per Month
2 - 250,000	$0.003 USD/NLU Item
250,001 - 5,000,000	$0.001 USD/NLU Item
5,000,000+	$0.0002 USD/NLU Item

Premium Tier1

Usage and Training Data is Private + Stored in an Isolated Single Tenant Environment

Transaction logging for service improvement is disabled by default

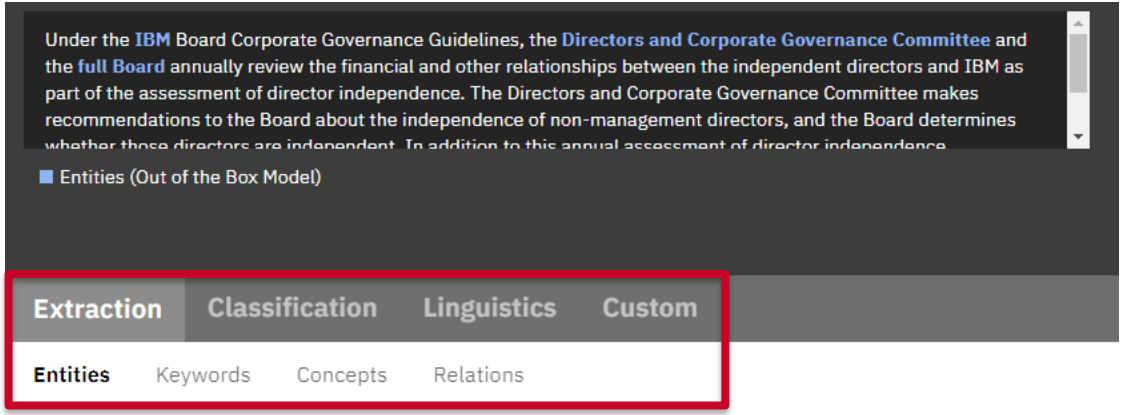
High Availability and Service Level Agreements on Uptime

IBM Cloud Service Endpoints

For more info or to purchase a premium plan, contact us at https://ibm.biz/contact-wdc-premium

Natural Language Understanding

- There are 9 features one can extract from a text using NLU
  - Entities
  - Keywords
  - Concepts
  - Relations
  - Sentiment
  - Emotion
  - Categories
  - Semantic Roles
  - Syntax



MIAMI UNIVERSITY

Natural Language Understanding

- API (https://cloud.ibm.com/apidocs/natural-language-understanding)
  - Request: REST protocol
    - GET operation
  - Response: JSON file
  - The API's documentation shows examples using *curl* and a specific Python module
    - We shall instead use the requests module we are already familiar with

Natural Language Understanding

- We will be using the same API keys today
  - You will learn how to get your own API keys in our next class
  - Please, do not overuse the provided API keys

```
base_url  = "https://api.us-south.natural-language-understanding.watson.cloud.ibm.com/instances/d6058b89-d39d-464c-a756-50658dd3124b"

api_key    = "LeNnOkIAOfA5VBWG6B7luAFzEJn4Q-z24AqSrzHaAGuG"
```

Natural Language Understanding

- Let's use the following text as example

```
content  =  "She's got a smile that it seems to me
                reminds me of childhood memories
                where everything was as fresh as the bright blue sky"
```

- We are going to send this text to IBM Cloud using a REST request
  - Note how there is no preprocessing

Natural Language Understanding

- Retrieving sentiment
  - REST request: GET operation with authentication
    - How do I know this? From IBM Cloud's documentation

```
key_values = {'version': '2021-08-01', 'text': content, 'features':'sentiment'}
response = requests.get(base_url+"/v1/analyze", key_values, auth = ('apikey', api_key))

response.json()
```

Natural Language Understanding

- Retrieving sentiment, keywords, emotion, and concepts
  - Make sure that there is no blank space in between the “features”

```
key_values = {'version': '2021-08-01', 'text': content, 'features':'sentiment,emotion,keywords'}
response = requests.get(base_url+"/v1/analyze", key_values, auth = ('apikey', api_key))

response.json()
```

Natural Language Understanding

- It is relatively easy to use cloud services to perform sentiment analysis
  - Arguably easier than what we did before using the bag-of-words approach
- Google, Microsoft, and Amazon, also offer similar services
  - How good are these services?
    - Search for the paper "*Off-the-Shelf Technologies for Sentiment Analysis of Social Media Data: Two Empirical Studies*" written by a previous ISA 414 student
  - How do they work?
    - Who knows … (proprietary technologies)

Summary

- We learned about different cloud services
  - IaaS, PaaS, SaaS, XaaS
- We learned about IBM Cloud
  - Can you think of interesting ways of applying this in your project?
- Interesting story: Netflix and the cloud
  - https://media.netflix.com/en/company-blog/completing-the-netflix-cloud-migration
- Next lecture: cloud computing and storage (part II)
  - Business considerations and hands-on time with some SaaS