CS4650 Topic 1: Introduction and Key Terms

Introduction

- In this class, we will talk about several related concepts:
  - Big Data -- dealing with data that is too big to fit on one computer
  - Analytics -- extracting meaningful information from this massive amount of data
  - Machine Learning -- techniques to help with analytics when we don't know where to look
  - Cloud Computing -- since this doesn't fit on one computer, how to harness many computers
- In today's topic, we will define many of the key terms used in these four areas.
- This will provide a framework. The remaining lectures in this class will fill out, populate the framework!

Big Data

Big Data

- What is Big Data?
  - Data that cannot be stored on one computer.
    - Consequently the data must be distributed among multiple computers.
  - Data in such quantity that it is virtually impossible to process using conventional techniques.
    - Consequently the data must be processed using distributed algorithms.
  - Massive amounts of structured and unstructured data that may come from a wide variety of sources.
- Many of the things we will learn about Big Data can also be applied to Small Data, but many of the things you can do with Small Data cannot be done, at least in the same way, with Big Data.

Structured vs Unstructured Data

- Structured data:
  - Clearly defined data types that is stored in a predefined order.
  - Explicit relationships link various items in the data.
  - The data is easily sortable and maintainable.
  - Example: College's list of students, classes, grades.
- Unstructured data:
  - A conglomeration of varied types of data, each stored in its native format.
  - No explicit links between entries.
  - Data is much more difficult to collect, process, and analyze.
  - Example: Logs from web servers or from automated weather stations.

Data Warehouse, Data Lake

- *Database* -- The files and programs that hold the data for one particular set or collection.
  - A school has a database that holds the student records. The class records may be considered part of the same database, or may have its own database.
  - The files that hold the records from the automated weather station are a database. If there are a collection of weather stations, they might each have their own database, or all of the records collectively may be considered one database.
- A *Data Warehouse* is a central repository holding one or more databases.
  - Since we are talking *big data*, this central repository will not be a single computer or single file system, but most likely consists of a cluster of such devices. But conceptually, users access one data warehouse, which happens to be internally constructed as a cluster.
- Some people consider a *Data Warehouse* holds structured data and a *Data Lake* holds unstructured data.

Accessing Databases

- Given a database, there are four basic tasks that can be performed:
  - *Query*: Copy some data from the database. *Search*, *Fetch*, *Get, Read, Retrieve*.
  - *Insert*: Add new data to the database. *Add, Post, Store, Write, Create*.
  - *Update*: Change existing data. *Put, Replace, Change, Modify.*
  - *Delete*: Remove some data.
- Databases do not necessarily implement all four of these actions.
- For example, many big data databases only allow for insert and query.

Accessing Databases

- In most cases, a database *sits outside* of an application.
- The database has management software that implements the operations of the database by accessing files which store the data.
- The application does not directly access the files of the database, but rather send commands to the management software.
- The combination of management software and data files, the *database server*, is a *service* which can be used by one or more application.
- The database server 'implements' an Application Programming Interface (API), a set of instructions by which the application directs the server to perform actions.

Accessing Databases: APIs

- The industry standard API for interacting with a structured database is SQL – Structured Query Language.
  - SQL can be considered a programming language for expressing the actions to be implemented by the database.
- Unstructured data does not have a standard API, primarily due to the loose structure of the data. Instead, multiple ad-hoc APIs exist.
- The various unstructured database systems and APIs are called *NoSQL*, which means "not-only-SQL".

Data Quality

- Just because you have a lot of data does not mean that the data is useful.
- Sometimes you are looking for a needle in a haystack, or a diamond in the rough.
- Some data is just noise.
- A measure of *data quality* has been developed. The goal is to collect data of high quality.
- There is a danger, however: From one person's perspective, much of the data might be useless, but from another perspective, it might contain valuable information. One person's flower might be another person's weed.

Data Quality

- Data Quality is a measure of the suitability of the data for a particular use, rated using the following four criteria:
  - *Accuracy* -- Does the data correctly represent real world values?
  - *Completeness* -- The degree to which all values are filled or all values are represented.
  - *Consistency* -- The degree to which similar or related data values align.
  - *Conformity* -- The degree to which the data values align with the business goals.

Data Cleansing

- Some data sets might have a relatively low data quality, but through *data cleansing*, the quality can be improved.
- Examples of data cleansing:
  - Elimination of duplicate records
  - Resolve missing values, NULL values, NaN (Not-a-Number) values
  - Converting values from inconsistent or archaic representations to more useful formats
  - Basically, fixing errors causing low scores in each of the four previously mentioned criteria
- We are talking big data, so these fixes are not done by hand! Instead:
  - Writing custom programs
  - Using 3rd party tools.

Internet of Things

- Starting in the 1990s, many devices have been 'connected' to the Internet that have embedded computers. This has been called IoT, the Internet of Things.
- One of the first devices that was humorously suggested was a toaster.
- There are many devices now available:
  - Sensors
  - Wearables
  - Robotics
  - Cars
  - Lights
  - Thermostats
  - Locks
  - Medical devices
- These devices tend to generate large quantities of data (Big Data).

Data Engineering

- We have talked about many aspects of Big Data.
- As the industry grows, new types of jobs are created.

- Data Engineering – A discipline that focuses on identification of data sources, collection, curation, and storage of data.

Analytics

Analytics

- Once we have this vast amount of data, what do we do with it?

- Data Science – The discipline of applying advanced analytic techniques to extract valuable information from data for decision making and strategic planning.
- Data Mining -- The process of identifying patterns that exist within large sets of data.
    - Many involve the use of statistics, data set queries, visualization tools, programming languages, and machine learning.

Business Intelligence

- The goal of Big Data, Analytics, Machine Learning is *business intelligence*.
- *Business Intelligence*: Use of tools (data mining, machine learning, and visualization) to convert data to actionable insights and recommendations.
  - Analyze and transform data
  - Extract valuable business insights
  - Enable decision-making.

Analytics

- Analytics – Examination of data to answer a question.
- Descriptive analytics – “What happened?”
  - This typically used data visualization.
- Diagnostic analytics – “Why did it happen?"
  - This uses data mining, statistics, and machine learning.
- Predictive analytics – “What is likely to happen?”
  - This uses machine learning and AI.
  - This is not an attempt to second-guess the future, but rather forecasting with probabilities.

Information Design

- Information Design – The practice of presenting information in a way that fosters efficient and effective understanding of that information.
- Data Visualization – Use of charts, graphs, and maps to represent data more clearly.
    - Which chart/graph to use is based on how to represent data most clearly.

Some Types of Analytics

- Regression – curve fitting
  - Can we deduce a formula that approximates the relationship between input and output.
- Data Clustering – Process of grouping related data set items in one or more clusters.
  - Clustering is unsupervised learning
  - There are many algorithms
  - Does not use training data set.
- Data Classification –Process of assigning data to groups or categories.
  - i.e. email: valid or spam
  - i.e. bank transaction: legitimate or fraud
- Data Association – Process of identifying key relationships between variables.
  - Example: shopping cart analysis
  - Examine how one item in shopping cart (antecedent) influences the addition of a second item (consequent)
  - This makes a recommendation

A Few More Terms

- Visual Programming – Use drag/drop objects in a workspace instead of programming
  - Used by non-technical workers/employees.
- Data Analyst – many are non-technical.
  - Use programs like Excel with 3rd party tools to do data mining operations.

Tools We Will Use

- Hadoop – An open-source software framework and collection of tools that allow for storage, retrieval, and analysis of very large data sets.
- MapReduce – A programming model that is used by many big data analytics tasks.
    - This works on large, unstructured data sets.
    - The Map portion extracts a subset of the information from each record
    - The Reduce portion combines the mapped data to generate the final report

Machine Learning

Machine Learning

- One of our books stressed, several times:

  "Computers cannot think, they cannot understand"

- Computers can process a lot of information.
- Computers can be programmed with various types of programs to find various patterns in the data.
- Some of the programming techniques mimic some of the ways that the human brain works.
- The output of the program may be similar to the output that a human would have produced.

Artificial Intelligence (AI)

- Artificial Intelligence (AI) – The ability of machines to mimic the capabilities of the human mind, such as:
  - Learning from examples and experience
  - Recognizing objects
  - 'Understanding' and responding to language
  - Making decisions
  - Solving problems

Machine Learning

- Machine Learning:
  - A subset of AI
  - Provides systems the ability to automatically learn and improve from experience without being explicitly programmed
  - Use of data pattern recognition algorithms to solve problems
- There are two basic forms of Machine Learning:
  - Supervised: Examine training set data to learn to identify patterns
  - Unsupervised: Does not use a training set

Deep Learning

- Deep Learning:
  - A subset of Machine Learning
  - Based on artificial neural networks that are inspired by the structure of the human brain
  - Learns from vast amount of data and is particularly good at finding patterns from unstructured data such as text and images.

Augmented Intelligence

- Augmented Intelligence -- Refers to human-centered partnership that brings people and AI together to enhance cognitive performance, including:
  - Learning
  - Decision making
  - New experiences.

Cloud Computing

Cloud Computing

- Cloud Computing – The delivery of IT services on-demand over shared networked computing resources.
- Initially, large centralized computers did the work, people shared these services through connected terminals.
- With the widespread adoption of personal computers, much of the work migrated to individual computers.
- Companies built data centers to share expensive servers and devices.
- The Internet allowed for people to access information from widely distributed sources.
- In addition to providing information, the Internet now provides for remote computations.

Cloud Computing

- In many ways, computing has become a utility, much like water, electricity, and gas.
  - Companies no longer generate their own electricity or find their own supply of water.
  - Instead, they purchase these from utility companies.
  - In the same say, companies do not have to pay for and provide upkeep for computing resources, they can purchase these from suppliers on the Internet.
    - For example, rather than buy a bunch of hard drives to store data, and then provide upkeep, maintenance, backups, etc, they can just purchase space from data storage providers.
    - If a company needs to perform a complex computation, rather than buying computers for this one-time use, they can 'rent' time on clusters of computers from a computation provider.

Cloud Computing

- The term 'cloud' has been used, and overused!, a lot.
- We will explore several versions of 'cloud'

Public Cloud

- Public Cloud – A cloud infrastructure system that is hosted by a cloud service provider and can be accessed by anyone with an Internet connection.
- Examples: Amazon’s EC2, IBM’s Blue Cloud, and Google Drive.
- Users don’t have to purchase software, hardware, or infrastructure. These are all managed by the cloud provider, who may or may not charge a fee.

Private, Hybrid Clouds

- Private Cloud – A cloud infrastructure system that is used by multiple users within one single entity (business, organization, etc).
- It can be operated by the home organization, a third party, or both.
- It can be located on-premises or off-site.
- Private clouds can provide better security than a public cloud.

- Hybrid Cloud – A setup which is a combination of both public and private clouds. Some portions or aspects are private, others are public.

Internal, External Clouds

Internal Cloud – A private cloud service created or offered by an internal IT department, strictly for in-house use.

External Cloud – This is a cloud service where fees are most likely charged, and offers customization to suit clients' needs. The cloud may be either private or public.

More Clouds

Personal Cloud – Rebranding, giving an old technology a new name! This is a network-attached-storage (NAS) device. A computer connected to a network for purposes of a dedicated data storage.

Consumer Cloud – Cloud offerings mainly intended for people for their personal use, such as DropBox.

Vertical Cloud – A cloud environment that it built to serve the needs across several specific industries, such as healthcare, financial services, etc.

- Infrastructure – Catchall term describing all IT resources, both virtual and hardware, that support a given IT environment.
- Middleware – Software that serves as a bridge between applications and components.
- Data Migration – Moving data between multiple formats, servers, storage systems, or warehouses.
- Cloud Backup – Backing up data to a remote, cloud-based server. The data is stored in and accessed from a network of interconnected resources. Dropbox, One-Drive. Carbonite.

Virtual Machines

- Virtual Machine – A VM is software that emulates a computer, used for running an operating system or applications.
  - The physical computer typically has a host operating system (the 'real' operating system for the computer).
  - On that computer is a program called the *hypervisor*, which acts like the hardware for a different computer.
  - 'On top' of the hypervisor is another operating system, the guest OS. This is the *virtual machine*.
  - Programs running on the guest OS 'think' they are running on one type of computer, but actually are running on a different computer.
  - One physical machine might actually be running several virtual machines.
- Virtual Desktop Infrastructure – Virtualization technology that hosts a desktop operating system on a virtual machine.

Cloud Services

- A company may develop a product that it wants to make available as a 'cloud service', so customers can access this product over the Internet, from any connected browser.
- However, this company may not have the resources to host the actual computers and other devices required for their product to run on the cloud.
- Cloud providers, who do host the actual computers, offer various 'packages' which can help this company.
- These packages go by the name "*X*-as-a-Service".

Backend-as-a-Service

- Suppose the company *MobileAppsRUs* developed a mobile app, or a web app, which needs a cloud backend. Perhaps this backend provides file storage, or provides communication between on-line gaming devices.
- This cloud backend requires web-connected servers (computers), routers, storage devices, all of which require maintenance and support. A complex, costly undertaking.
- Suppose another company, *CloudsRUs*, has this infrastructure, and provides a package called BaaS (Backend-as-a-Service). This package supplies the tools and services that *MobileAppsRUs* can use to build their cloud backend.
- The mobile apps communicate with the *MobileAppsRUs* backend, which is running on the *CloudsRUs* computers.

SaaS (Software-as-a-Service)

- SaaS (Software-as-a-Service), represents software applications hosted by a vendor and sold as subscription licenses to users.
- Examples: Adobe’s Creative Cloud, Microsoft’s Office 365, and Google Docs (although Microsoft and Google are using their own clouds for these).
- SaaS comprises software applications which are run on distantly located computers that happen to be owned as well as operated by others.
- Key benefits: Instance access and usage of applications, accessibility from any connected machine, no likely loss of data.

PaaS (Platform-as-a-Service)

- PaaS (Platform-as-a-Service) a product where the cloud service provider offers users the necessary software and hardware for the creation, deployment, and management of applications, all via the Internet.
- This is possible without the developer of the application having to purchase hardware, software, management, and even hosting.
- Key benefits: applications may be deployed really fast, without worrying about the platform, while savings costs, management, upgrades.
- An example might be that rather than buying expensive, powerful workstations for employees, the company rents PaaS computers for their employees to use.

IaaS (Infrastructure-as-a-Service)

- IaaS (Infrastructure-as-a-Service) a virtual environment delivered to a customer by a cloud provider.
- This features networking equipment, servers, storage, and software.
- Customers can build their own XaaS products on top of this IaaS service.
- Key benefits are not having to invest in hardware or upgrades, and availability of dynamic and flexible services.

Cloud Providers

- Cloud Provider – An organization or business that offers access to cloud computing services, generally for a fee.
- AWS – Amazon Web Services, Amazon's cloud computing platform (the Internet's most-used platform)
- Azure – Microsoft's cloud computing platform. It provides IaaS and PaaS services.
- GCP – Google Cloud Platform, Google's entry into cloud computing. It offers IaaS and PaaS.
- Amazon EC2 (Elastic Cloud Compute) – Part of AWS, this is scalable cloud computing. It is used for the deployment of cloud-based applications on virtual servers that can be rented. Users can even opt to pay for the resources they use by the hour.
- Amazon Simple Storage Service (S3) – Also part of AWS that allows for the storage and backup of data on the cloud. It is highly scalable, unlimited archiving and backup options.

A Few More Terms

- Service License Agreement (SLA) – the agreement between a cloud service provider (CSP) and a customer, outlining contractural terms such as availability, level of service, and performance.
- Elasticity – A cloud storage system's capability for adapting to client's fluctuating workload demands.
- Consumption-Based – The cloud computing pricing model where consumers are charged for how much of the service they use, rather than a block of time.
- Pay-As-You-Go – a means of purchasing cloud services from a provider that requires no money down; Instead, it's based on consumption or via subscription.
- Load Balancing – The distribution of workloads over a series of resources, such as servers, to assure that no single server suffers a point of failure.

Yet More Terms

- Multi-Tenancy – The ability of a single platform to run and hold a process, application, or virtual machine for many users.
- Cloud Portability – The ability of an application and data to move from one cloud service provider to another cloud service provider.
- Cloud Enablement – Make a could computing environment by writing the necessary software, infrastructure, and applications.
- Cloudstorming – Assembling multiple cloude computing environments.
- Cloud Broker – An entity which is responsible for managing relationships between customers and various cloud service providers.
- Cluster Computing – Computing using a cluster of pooled resources and interconnected processors.

Conclusion

- Whew! We talked about a lot of terms!
- As this class progresses, we will revisit these terms (so you don't have to remember all of these details).
- Hopefully this gives you an idea of what we will be studying, and how this all plugs together.