Big Data and Data Mining

1. **Overview**

* How Big Data is Driving Digital Transformation
* Data Science Skills and Big Data
* Introduction to Cloud
* Cloud for Data Science
* Foundations of Big Data
* Data Scientists at New York University
* What is Hadoop
* Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
* Data Mining

1. **How Big Data is Driving Digital Transformation**

Digital Transformation:

* Affects business operations. Updating existing processes and operations. And creating new ones to harness the benefits of new technologies.
* Integrate digital technology into all areas of an organization that results in fundamental changes to how it operates and delivers value to customers.
* It is an organizational and cultural change driven by Data Science and especially Big Data.

Vast amount of data + Advanced Analysis => Digital Transformation

Digital Transformation is not simply duplicating existing processes in digital form; the in-depth analysis of how the business operates helps organizations discover how to improve their processes and operations.

1. **Introduction to Cloud**

Cloud computing:

* Delivery of on-demand computing resources (Networks, Servers, Storage, Application, Services, Data centers) Over the Internet on a pay-for-use basis
* Applications and data that users access over the internet rather than locally:
  + Online web apps
  + Secure online business applications
  + Storing personal files
    - Google Drive
    - OneDrive
    - Dropbox

Benefits:

* No need to purchase applications and install them on local computer
* Use online versions of applications and pay monthly subscription
* More cost-effective
* Access most current software versions
* Save local storage space
* Working collaboratively in real time

Cloud computing:

* Five characteristics
  + On-demand self-service: access processing powers, storage, network via a simple UI
  + Broad network access: can be accessed through many platforms and devices (Mobile phones, Tablet, Laptops, Workstations, …)
  + Resource pooling: gives provider economy of scale. Making Cloud cost-efficient. Using a multi-tenant, resources are pooled to serve multiple customers and resources are dynamically assigned and reassigned according to demand.
  + Rapid elasticity: you can access more resources when you need them and scale back when you don’t.
  + Measured service: you only pay for what you use or reserve as you go. Resorce usage is monitored and measured based on consumer utilization.
* Three development models
  + Public: owned by a cloud provider, but its usage is shared by other companies
  + Private: cloud infrastructure is provisioned for exclusive use by a single organization.
  + Hybrid: has both
* Three service models
  + Infrastructure as a Service (IaaS): you can access the infrastructure and physical computing resources such as servers, networking, storage, and data center space
  + Platform as a Service (PaaS): you can access the platform that comprises the hardware and software tools that are usually needed to develop and deploy applications to users over the internet
  + Software as a Service (SaaS): software licensing and delivery model in which software and applications are centrally hosted and licensed on a subscription basis.

1. **Cloud for Data Science**

Cloud enables instant access to Open source technologies (Apache spark) without the need to install and configure them locally. Using the cloud gives you access to up-to-date tools and libraries without the worry of maintaining them and ensuring that they are up to date.

The cloud is accessible from everywhere and in every time zone, enabling multiple collaborators and teams to access the data simultaneously and work together on producing a solution.

Some big tech companies offer cloud platforms, allowing you to become familiar in a pre-built environment:

* IBM: IBM Cloud
  + IBM provides Skills Network Labs to learners registered at any of the learning portals on the IBM Developer Skills Network
  + Access to tools like Jupiter notebooks and Spark clusters => Create your own data science project and solutions.
* Amazon: AWS
* Google: Google Cloud
* With practice and familiarity, you will discover how the Cloud dramatically enhances productivity for data scientists.

1. **Foundation of Big Data**

A white text on a white background

Description automatically generated

The V’s of Big Data:

* Velocity
* Volume
* Variety
* Veracity
* Value

1. **Data Scientists at New York University**

* Programming
* CS
* MBA

1. **What is Hadoop**
2. **Big Data Processing Tools: Hadoop, HDFS, Hive, and spark**

A yellow and black logo

Description automatically generated

A black background with white text

Description automatically generated

**Hadoop:**

* Distributed storage and processing of large datasets across clusters of computers (A Node: A computer; A Cluster: A Group of Computers)
* Each Node provides storage and computation.
* Hadoop provides a reliable, scalable, and cost-effective solution for storing data with no format requirements.
* **Benefits include:**
  + **Better real-time data-driven decisions**: Incorporates emerging data formats not traditionally used in data warehouses.
  + **Improved data access and analysis:** Provides real-time, self-service access to stakeholders.
  + **Data offload and consolidation:** Optimizes and streamlines costs by consolidating data, including cold data (least frequently used data), across the organization.
* **Hadoop Distributed File System (HDFS):** Is a storage system for big data that runs on multiple commodity hardware connected through a network.
  + Provides scalable and reliable big data storage by partitioning files over multiple nodes.
  + Splits large files across multiple computers, allowing parallel access to them.
  + Replicates file blocks on different nodes to prevent data loss
  + Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover.
  + Access to streaming data, because HDFS supports high data throughput rates.
  + Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
  + Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems

A black background with white text

Description automatically generated

Hive:

* Is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase
* Queries have high frequency -> not suitable for applications that need fast response times
* Read-based -> not suitable for transaction processing that involves a high percentage of write operations
* Hive is better suited for:
  + Data warehousing tasks such as ETL (Extract, Transform, Load), reporting, and data analysis
  + Easy access to data via SQL

A black background with white text

Description automatically generated

Spark:

* Spark is a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications:
  + Interactive Analytics
  + Streams Processing
  + Machine Learning
  + Data Integration
  + ETL
* Key attributes:
  + Has in-memory processing which significantly increases speed of computations
  + Provides interfaces for major programming languages such as Java, Scala, Python, R and SQL
  + Can run using its standalone clustering technology.
  + Can also run on top of other infrastructure, such as Hadoop
  + Can access data in a large variety of data sources, including HDFS and Hive.
  + Processes streaming data fast.
  + Performs complex analytics in real-time

1. **Data Mining**

[Link](https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTAxRU4tU2tpbGxzTmV0d29yay9sYWJzL01vZHVsZSUyMDIvRGF0YV9NaW5pbmcubWQiLCJ0b29sX3R5cGUiOiJpbnN0cnVjdGlvbmFsLWxhYiIsImFkbWluIjpmYWxzZSwiaWF0IjoxNzExNTYyNjQ2fQ.5JvPCUY21HyYZZbA54EtM7EWZph9cxv5vpGAz7yFul0)

Data mining steps:

1/ **Set up goals:** Identify the key questions that need to be answered and consider cost-benefit trade-offs.

2/ **Selecting data:** Identify sources of data or plan new data collection initiatives, including surveys. Identify the right kind of data.

3/ **Preprocessing Data:** Exclude, replace, or develop a specific formal method to deal with messy, erroneous, irrelevant, missing data. Determine whether the data is missing randomly or systematically.

4/ **Transforming Data:** Determine appropriate format in which data must be stored, reduce attributes without significant loss in information. Variables may need to be transformed to help explain the phenomenon being studied.

5/ **Storing Data:** Transformed data must be stored in a format for easy data mining. Data reading/writing should be unrestricted and immediate. Data must be stored securely.

6/ **Mining Data:** This step covers data analysis methods (parametric and non-parametric methods), and machine-learning algorithms. Data visualization is a good starting point to develop understanding of the trends hidden in the data set.

7/ **Evaluating Mining Results:** Do a formal evaluation of the results. Formal evaluation may include "in-sample forecast": testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. The results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

1. **Summary**
2. **Glossary**

[Link](https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTVNraWxsc05ldHdvcmstRFMwMTAxRU4tQ291cnNlcmEvbGFicy8yMDA0NTcuMDY5X00yX0wxX0dsb3NzYXJ5Lm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTcxMTU2MjY1MX0.SQgRMP4CX_d6NJ9ERKW-XqJk71UzXWsgikcLtk6Qmdo)