Data Engineering in the Cloud

Introduction

Xuemao Zhang
East Stroudsburg University

January 18, 2025

What's covered in this lecture?

- Math 418 Course Outline
  - Course Objectives
  - Tentative Contents
  - Assessments
  - References
- What is Big Data?
- Introduction to Data Engineering
- Introduction to Cloud Computing

Math 402 Course Outline

Course Admin Information

- Instructor: Dr. Xuemao Zhang
    - Office: SciTech Rm 128
    - Email: xzhang2@esu.edu
- Lecture Hours:
    - MWF: 10:00–10:50Am
- Department Secretary: Christine Getz
    - Office: SciTech Rm 118
    - Email: cgetz@esu.edu
    - Telephone: 570-422-3447

Course Objectives

- This course will cover the knowledge and skills to design, build, and manage scalable data engineering solutions using cloud-based technologies.
- **Platform:** Microsoft Azure
- **Programming:** SQL, Python and Scala
- **You will learn:**
  - Utilizing Microsoft Azure for data storage, processing, and analysis
  - Implementing data pipelines and ETL (Extract, Transform and Load) processes in a cloud environment
  - Techniques for optimizing data workflows and ensuring data integrity
  - Design, build, and manage scalable data engineering solutions in the cloud

Tentative Contentes

- Introduction to Data Engineering
- Introduction to Microsoft Azure
- Basic SQL Queries for Data Engineering
- Using Scala in Big Data Processing
- Azure Data Engineering
  - Azure Blob Storage
  - Azure Data Lake
  - Azure SQL Database and Cosmos DB
  - Azure Synapse Analytics
  - Azure Databricks
  - Azure Data Factory; Building ETL Pipelines
  - Apache Spark on Azure
  - Azure stream analytics
  - Capstone Project

Assessments

- 8% Class attendance
- 24% in-class quizzes (4 sets) on SQL, Scala and Azure fundamentals
- 48% projects (8 sets)
- 20% Final project, consisting of
  - Oral presentation: 5%
  - Coding and Written report: 15%

References

- Nagaraj Venkatesan and Ahmad Osama (2022). *Azure Data Engineering Cookbook* . https://www.packtpub.com/product/azure-data-engineering-cookbook-second-edition/9781803246789
- *W3Schools Online Web Tutorials*. https://www.w3schools.com/
- *Tutorialspoint*. https://www.tutorialspoint.com/microsoft_azure/index.htm
- *Geeksforgeeks*. https://www.geeksforgeeks.org/microsoft-azure/

What is Big Data?

- Big Data is a term used to describe a massive volume of data which is large and complex that it becomes difficult or impossible to store and process using traditional data processing systems.
- One traditional processing system is RDBMS (Relational Database Management System).
- The concept of Big Data is often characterized by the “3Vs”:
  - Volume: Big Data involves terabytes, petabytes, or even exabytes of data.
  - Velocity: The speed at which data is generated, processed, and analyzed.
  - Variety: The different types of data, including structured, semi-structured, and unstructured data: csv, JSON, XML, Parquet

Classification of Big Data

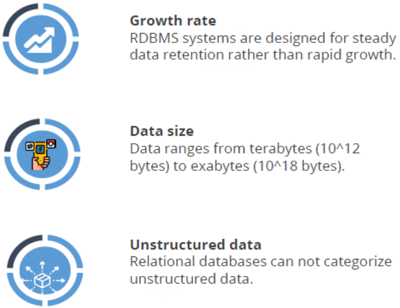
- Structured data: Data that is highly organized and easily searchable within fixed fields in a database or spreadsheet.
  - CSV, Excel, and SQL databases
- Semi-structured data: Data that does not conform to a rigid structure but still contains tags or markers to separate data elements.
  - JSON (JavaScript Object Notation): Data format used for APIs and web services.
  - XML (eXtensible Markup Language): Data format used for documents and data transfer.
  - Email: Contains structured fields like sender, recipient, and date, but the body of the email is unstructured.
  - logs
- Unstructured data: Data that lacks a predefined format or structure, making it more challenging to collect, process, and analyze. It is typically text-heavy or multimedia content.
  - Image, Audio, Video
  - Word files, PDFs, and other textual data.

Challenges of Big Data

- Storage
- Data processing

What is Big Data?

- Drawbacks of RDBMS (Relational Database Management System)
  - Limited Scalability: Vertical scaling
  - Schema Rigidity: They cannot accommodate unstructured/semi-structured data
  - Limited Ingestion Speed: They are not for high velocity data



Growth rate
RDBMS systems are designed for steady data retention rather than rapid growth.

Data size
Data ranges from terabytes (10^12 bytes) to exabytes (10^18 bytes).

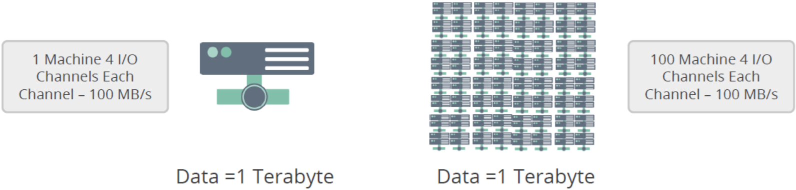
Unstructured data
Relational databases can not categorize unstructured data.

What is Big Data?

- RDBMS is not parallel programming

Distributed systems for Big Data

- A distributed system is a collection of **independent** computers or nodes that are linked together using network.
- These computers communicate and coordinate their actions by passing messages to one another, effectively sharing resources, and collaborating to solve complex problems.



1 Machine 4 I/O Channels Each Channel – 100 MB/s

Data =1 Terabyte

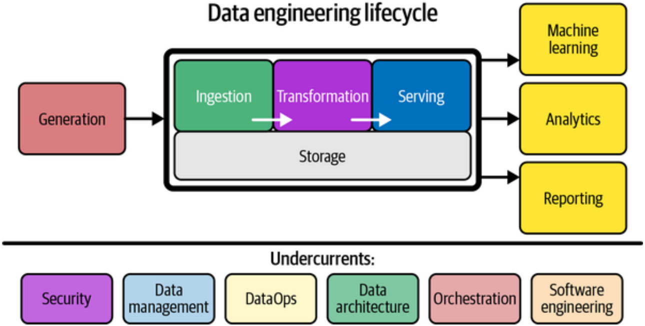
Data =1 Terabyte

100 Machine 4 I/O Channels Each Channel – 100 MB/s

Introduction to Data Engineering

- Data engineering involves designing, building, and managing systems for collecting, storing, and analyzing data at scale.
- Definition from *Data Engineering and Its Main Concepts* by AlexSoft:
  - Data engineering is a set of operations aimed at creating interfaces and mechanisms for the flow and access of information. It takes dedicated specialists—data engineers—to maintain data so that it remains available and usable by others. In short, data engineers set up and operate the organization's data infrastructure, preparing it for further analysis by data analysts and scientists.

The Data Engineering Lifecycle



Skills a Data Engineer Needs

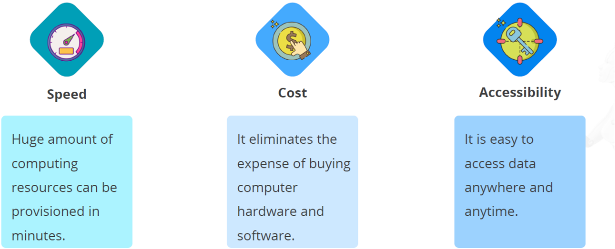
- Programming Languages
  - SQL
  - Python
  - Scala: Important for working with big data frameworks like Apache Spark.
- Data Processing
  - Batch Processing: Apache Spark
  - Stream Processing: Apache Kafka
- Big Data Technologies
  - Hadoop Ecosystem: Understanding of Hadoop components like HDFS, YARN, and MapReduce.
    - Spark Ecosystem: Skills in Spark Core, Spark SQL, Spark Streaming, and MLlib.
- ETL (Extract, Transform, Load) Processes
- Cloud Platforms

Introduction to Cloud Computing

- Cloud computing is accessing or storing the data over the Internet. It refers to the delivery of computing services like servers, database, storage, networking, analytics, and visualization.

Introduction to Cloud Computing

- Advantages of Cloud Computing



Speed

Huge amount of computing resources can be provisioned in minutes.

Cost

It eliminates the expense of buying computer hardware and software.

Accessibility

It is easy to access data anywhere and anytime.

Introduction to Cloud Computing

- Advantages of Cloud Computing

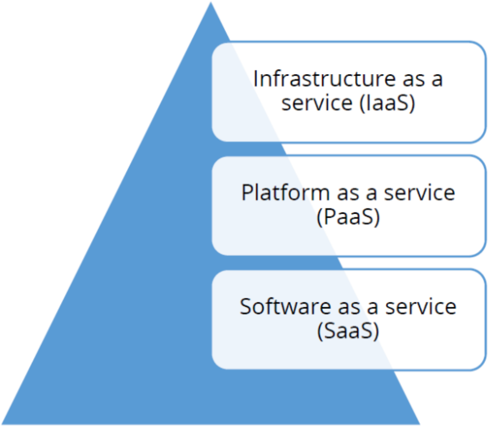
Scalability

Easy to scale up your cloud capacity based on the requirement.

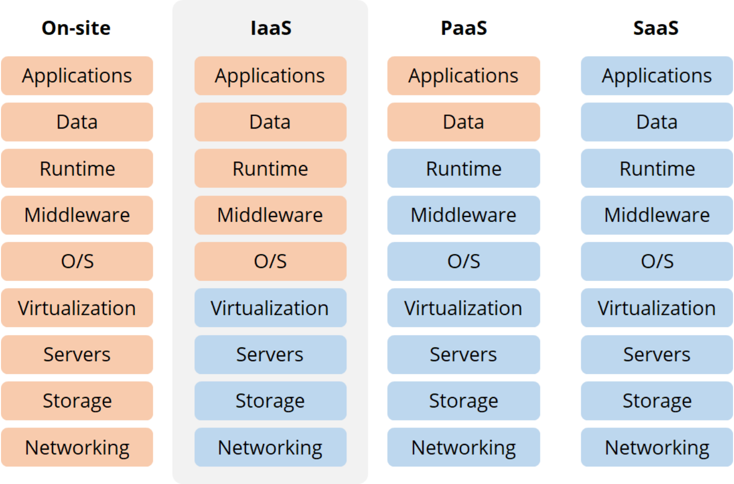
Security

Your data is stored in centralized and secure location.

Types of Cloud Computing


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

Types of Cloud Computing

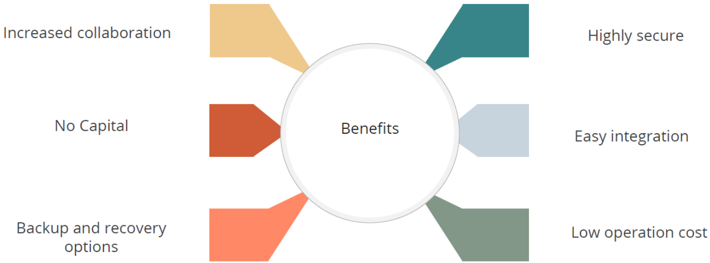


On-site	IaaS	PaaS	SaaS
Applications	Applications	Applications	Applications
Data	Data	Data	Data
Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware
O/S	O/S	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

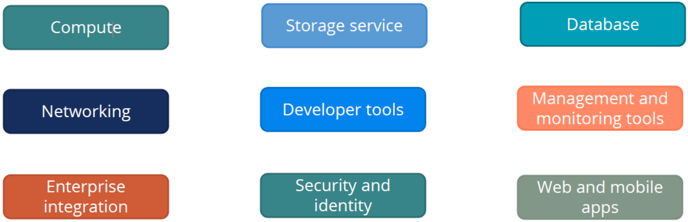
Microsoft Azure

- Microsoft Azure is a cloud computing platform that allows developers and IT professionals to design,deploy, and manage applications using Microsoft's global network of data centers.
    - For example, we will need a significant amount of money, time, and physical space to build a large server. Microsoft Azure comes to our aid in such circumstances.

Microsoft Azure



Service Domains in Azure



Compute

Storage service

Database

Networking

Developer tools

Management and monitoring tools

Enterprise integration

Security and identity

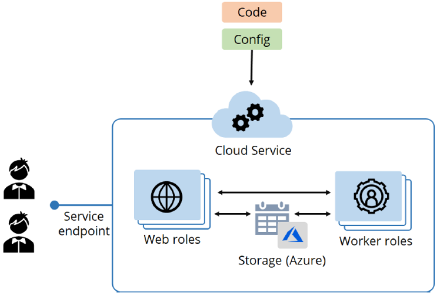
Web and mobile apps

Service Domains in Azure

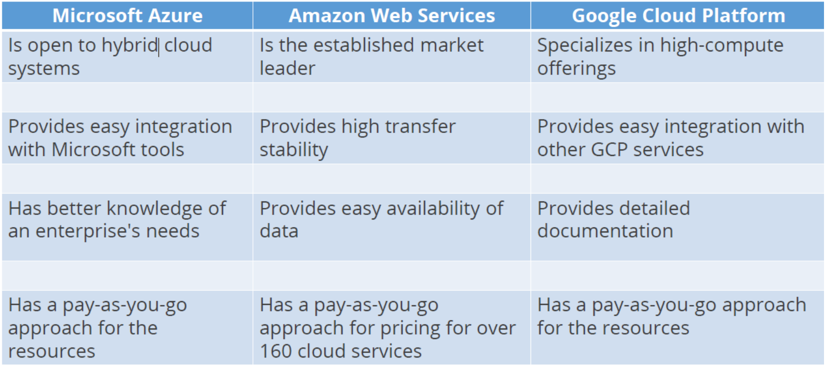
- Azure cloud services exemplify Platform as a service (PaaS) cloud computing.
- Azure cloud services are hosted on virtual machines in the same manner as App Service.

Azure Cloud Service Roles

- In Microsoft Azure, there are two primary roles within Cloud Services that help manage and execute applications: Web Roles and Worker Roles.
- Web Role: A Web Role is designed to host applications through the Internet Information Services (IIS) platform.
  - It is tailored for applications that serve HTTP requests, such as websites and web APIs.
- Worker Role: A Worker Role is designed to perform background processing without the need for IIS.
  - It runs as a standalone service that executes tasks and processes data independently.



Azure vs. AWS vs. GCP



Microsoft Azure	Amazon Web Services	Google Cloud Platform
Is open to hybrid cloud systems	Is the established market leader	Specializes in high-compute offerings
Provides easy integration with Microsoft tools	Provides high transfer stability	Provides easy integration with other GCP services
Has better knowledge of an enterprise's needs	Provides easy availability of data	Provides detailed documentation
Has a pay-as-you-go approach for the resources	Has a pay-as-you-go approach for pricing for over 160 cloud services	Has a pay-as-you-go approach for the resources

License