

# MapReduce II and Object Storage

The background of the slide features a photograph of the Alma Mater statue at the University of Illinois, which is a central figure in a long, flowing gown with her arms outstretched. The entire image is overlaid with a semi-transparent orange color, creating a monochromatic effect.

**CS 240 - The University of Illinois**

Wade Fagen-Ulmschneider

October 26, 2021

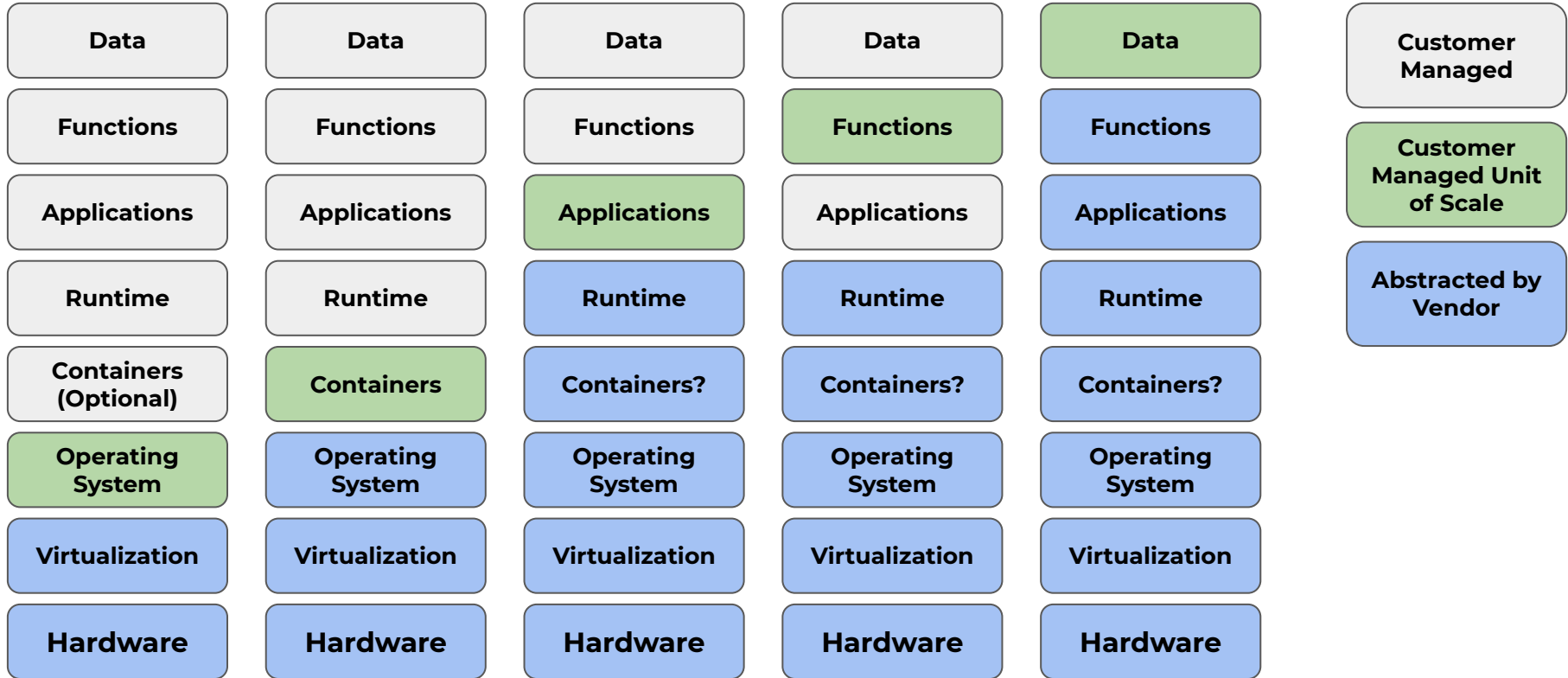
# IaaS

# CaaS

# PaaS

# FaaS

# SaaS



# MapReduce

A photograph of a crowd of people gathered around a statue of Alma Mater, overlaid with a semi-transparent orange filter. The text "MapReduce" is centered in white. The statue is the central focus, with people in the foreground and background. The background shows a large tree with bare branches. The overall scene is a public gathering, likely at a university.

# MapReduce

MapReduce is a **framework** for processing data that can be “**automatically parallelized**” and therefore scale massively.

# Apache Hadoop

*“The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.”*

-- <https://hadoop.apache.org/>

# Apache Spark

*“The most widely-used engine for scalable computing*

*Thousands of companies, including 80% of the Fortune 500, use Apache Spark. Over 2,000 contributors to the open source project from industry and academia.”*

-- <https://spark.apache.org/>

# Apache Hive

*“The Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive.”*

-- <https://hive.apache.org/>

# Cloud Providers



# Cloud Providers

AWS EMR:

<https://aws.amazon.com/emr/>

Azure HDInsight:

<https://azure.microsoft.com/en-us/services/hdinsight/>

Google DataFlow:

<https://cloud.google.com/dataflow>

# MP9 Design

# Final Project

A photograph of a crowd gathered around a statue of Alma Mater, overlaid with a semi-transparent orange filter. The text "Final Project" is centered in white. The background shows a large group of people, mostly young men, looking towards the statue. The statue is a large, standing figure in a long, flowing robe, with arms outstretched. The base of the statue has the words "ALMA MATER" visible. The overall scene is outdoors, with trees and foliage visible in the background.



# Data Storage



<b>Data Stores</b>	<b>Big Data / Data Pipelines</b>	<b>Object Storage</b>
<p>Useful for retrieving data for user requests (ms response times).</p> <p>Ex: User data, application data, etc</p>	<p>Useful for processing petabyte-scale datasets quickly to generate data summaries.</p>	<p>Useful for static files that do not change on a per-user request frequency.</p> <p>Ex: profile photo, images, data downloads, etc</p>

# Local File Storage

**/**

**/usr/**

**/usr/name/**

**/usr/name/Desktop/**

**C:/**

**C:/Users**

**C:/Users/name**

**C:/Users/name/Desktop/**

# Cloud Object Storage Systems

All objects are organized into \_\_\_\_\_:

- [Namespace]:
  
- [ACL]:



# Cloud Object Storage Systems

Each individual file is stored as an object, with attributes:

[Name]:

[Optional Tags]:

# Cloud Object Storage Systems

AWS S3

<https://aws.amazon.com/s3/>

Azure Blob Storage

<https://azure.microsoft.com/en-us/services/storage/blobs/>

Google Cloud Storage

<https://cloud.google.com/storage>