CS 240
Computer Systems

#18: MapReduce II and Object Storage

Oct. 28, 2021 · Wade Fagen-Ulmschneider

Overview of Technology Abstractions

IaaS Infrastructure as a Service	**CaaS** Containers as a Service	**PaaS** Platform as a Service	**FaaS** Functions as a Service	**SaaS** Software as a Service
Data	Data	Data	Data	Data
Functions	Functions	Functions	Functions	Functions
Applications	Applications	Applications	Applications	Applications
Runtime	Runtime	Runtime	Runtime	Runtime
Containers*	Containers	Containers*	Containers*	Containers*
OS	OS	OS	OS	OS
Virtualization	Virtualization	Virtualization	Virtualization	Virtualization
Hardware	Hardware	Hardware	Hardware	Hardware

	Abstracted by Provider/Vendor
	Customer Managed Unit of Scale
	Customer Managed

What labels of abstraction?

(1): Setting up and running Python on your CS 240 virtual machine?

(2): Using Docker to run `mongodb`?

(3): Connecting to a mongodb instance hosted on **`cs240-adm.cs`**?

(4): Using Docker to run a microservice that we provide for you?

(5): Using Queue@Illinois as a user for office hours in CS 240?

(6): Using github.com for MP submissions in CS 240?

MapReduce:

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

There are many pieces of a MapReduce framework:

- **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:**
  Similar to Hadoop, but does all processing in RAM instead of on a file system. Can be "100x faster on small workloads".

  "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 of MapReduce framework technologies:

- AWS EMR:
  https://aws.amazon.com/emr/
- Azure HDInsight:
  https://azure.microsoft.com/en-us/services/hdinsight/
- Google DataFlow:
  https://cloud.google.com/dataflow

Use Cases for MapReduce

MP9:
We're going to focus on the MapReduce paradigm itself -- allowing you to build `map` and `reduce` functions that run in separate processes. With minimal configuration, this code could run in with pyspark -- the Python Spark library to run your code on Apache Spark.

Final Project: Class Preferences!
- Cloud System? Course-wide Graph? Course-wide game?

Data Storage in the Cloud

Data Stores	Big Data / Data Pipelines	Object Storage
- Key-Value Stores - NoSQL Databases - SQL Databases - Domain-Purpose Databases	- MapReduce - Apache Hadoop - Apache Spark - Hive / Storm / etc	
Useful for retrieving data for user requests (ms response times). (Ex: User data, application data, etc)	Useful for processing petabyte-scale datasets quickly to generate data summaries.	Useful for static files that do not change on a per-user request frequency. (Ex: profile photo, images, data downloads, etc)

Local File System
Modern operating systems organize data into a hierarchical tree structure:

- `/` `C:/`
- `/usr/` `C:/Users`
- `/usr/name/` `C:/Users/name`
- `/usr/name/Desktop/` `C:/Users/name/Desktop/`

Unfortunately, files on a local system can only be accessed on that local system. For cloud-scale applications, how do we store files?

Object Cloud Storage Systems
All objects in cloud object storage are organized into _____. Each one has:

- [Namespace]:

- [ACL]:

Then, any number of files can be stored within a bucket. Each individual file has attributes:

- [Name]:

- [Optional Tags]:

Cloud Providers of MapReduce framework technologies:
- AWS S3
- Azure Blob Storage
- Google Storage