MapReduce II and Object Storage

CS 240 - The University of Illinois
Wade Fagen-Ulmschneider
October 26, 2021
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
Final Project
Data Storage
<table>
<thead>
<tr>
<th>Data Stores</th>
<th>Big Data / Data Pipelines</th>
<th>Object Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful for retrieving data for user requests (ms response times).</td>
<td>Useful for processing petabyte-scale datasets quickly to generate data summaries.</td>
<td>Useful for static files that do not change on a per-user request frequency.</td>
</tr>
<tr>
<td>Ex: User data, application data, etc</td>
<td></td>
<td>Ex: profile photo, images, data downloads, etc</td>
</tr>
</tbody>
</table>
Local File Storage

/                           C:(/
/usr/                      C:/Users
/usr/name/                 C:/Users/name
/usr/name/Desktop/        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