Data 604 – Data Management

Week 10 – Big Data & Cloud Computing

Jyotsna Potarazu

Data 604: Data Management

Topics:

- Big Data
- Hadoop Ecosystem
- Cloud Computing
- Price Estimates Cloud vs On Premise
- Homework 3 Assigned

Patty Delafuente, Adjunct Faculty

Learning Objectives

- Identify characteristics of Big Data and how it changed the database industry
- Identify characteristics of Data Lake
- Identify components of Hadoop
- Identify differences between Cloud and On-premise data services and solutions
- Be able to create and compare cost estimates for Cloud and On-premise services

The 5 Vs of Big Data

- Every minute:
  - more than 300,000 tweets are created
  - Netflix subscribers are streaming more than 70,000 hours of video at once
  - Apple users download 30,000 apps
  - Instagram users like almost two million photos
- Big Data encompasses both structured and highly unstructured forms of data

The 5 Vs of Big Data

- **Volume**: the amount of data, also referred to the data “at rest”
- **Velocity**: the speed at which data comes in and goes out, data “in motion”
- **Variety**: the range of data types and sources that are used, data in its “many forms”
- **Veracity**: the uncertainty of the data; data “in doubt”
- **Value**: TCO and ROI of the data

The 5 Vs of Big Data

- Examples:
  - Large-scale enterprise systems (e.g., ERP, CRM, SCM)
  - Social networks (e.g., Twitter, Weibo, WeChat)
  - Internet of Things
  - Open data

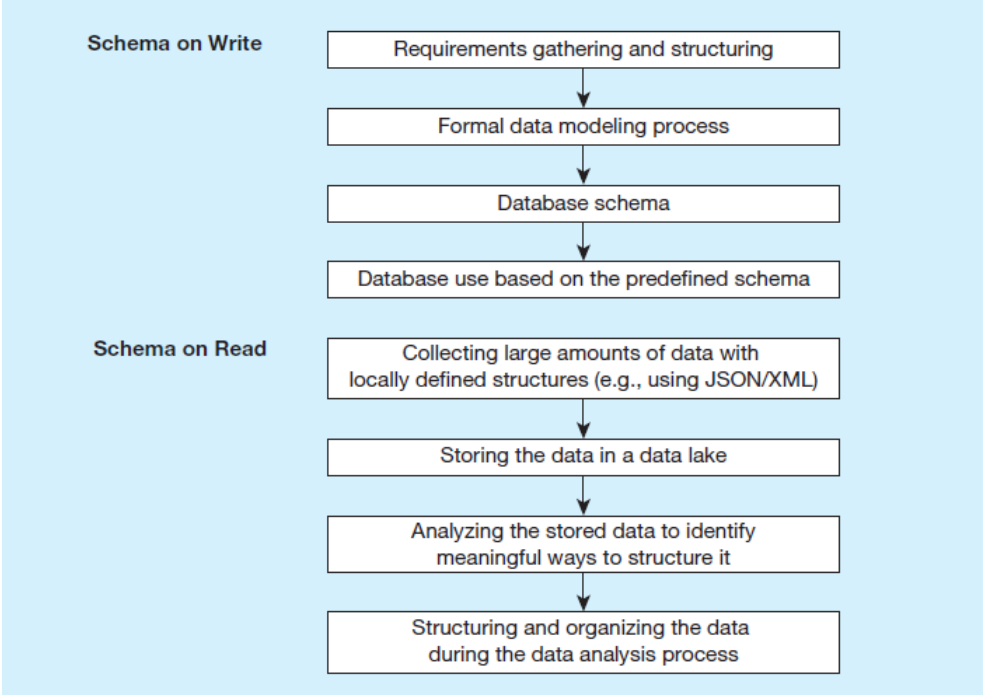
Big Data Storage

- Schema on Read, rather than Schema on Write
  - Schema on Write – preexisting data model, how traditional databases are designed (relational databases)
  - Schema on Read – data model determined later, depends on how you want to use it (XML, JSON)
  - Capture and store the data, and worry about how you want to use it later
- Data Lake
  - A large integrated repository for internal and external data that does not follow a predefined schema
  - Capture everything, dive in anywhere, flexible access

Schema on Write vs on Read

The big data approach

Traditional database design



Infrastructure Advances in Data Management

- Massively parallel processing (MPP)
  - divide a computing task (such as query processing) between multiple processors, speeding it up significantly
- In-memory DBMS
  - keep the entire database in primary memory, thus enabling significantly faster processing
- In-database analytics
  - no need to move large quantities of data to separate analytics tools for processing
- Columnar DBMS
  - reorient the data in the storage structures, leading to efficiencies in data warehousing and analytics applications

Hadoop

- Open-source software framework used for distributed storage and processing of big datasets
- Can be set up over a cluster of computers built from normal, commodity hardware
- Many vendors offer their implementation of a Hadoop stack (e.g., Amazon, Cloudera, Dell, Oracle, IBM, Microsoft)

Hadoop

- History of Hadoop
- The Hadoop stack

History of Hadoop

- Key building blocks:
  - Google File System: a file system that could be easily distributed across commodity hardware, while providing fault tolerance
  - Google MapReduce: a programming paradigm to write programs that can be automatically parallelized and executed across a cluster of different computers
- Nutch web crawler prototype developed by Doug Cutting
  - Later renamed to Hadoop
- In 2008, Yahoo! open-sourced Hadoop as “Apache Hadoop”

The Hadoop Stack

Four modules:

- Hadoop Common: a set of shared programming libraries used by the other modules
- Hadoop Distributed File System (HDFS): a Java-based file system to store data across multiple machines
- MapReduce framework: a programming model to process large sets of data in parallel
- YARN (Yet Another Resource Negotiator): handles the management and scheduling of resource requests in a distributed environment

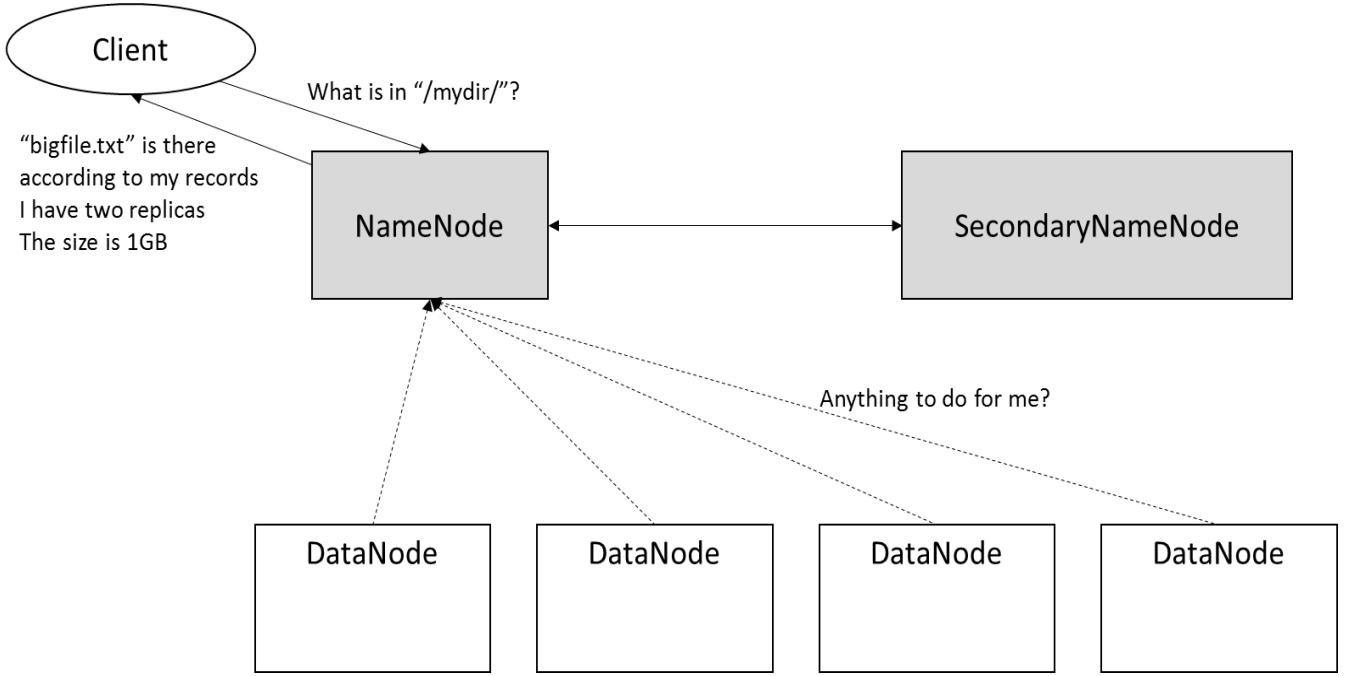
Hadoop Distributed File System (HDFS)

- Distributed file system to store data across a cluster of commodity machines
- High emphasis on fault tolerance
- HDFS cluster is composed of a NameNode and various DataNodes

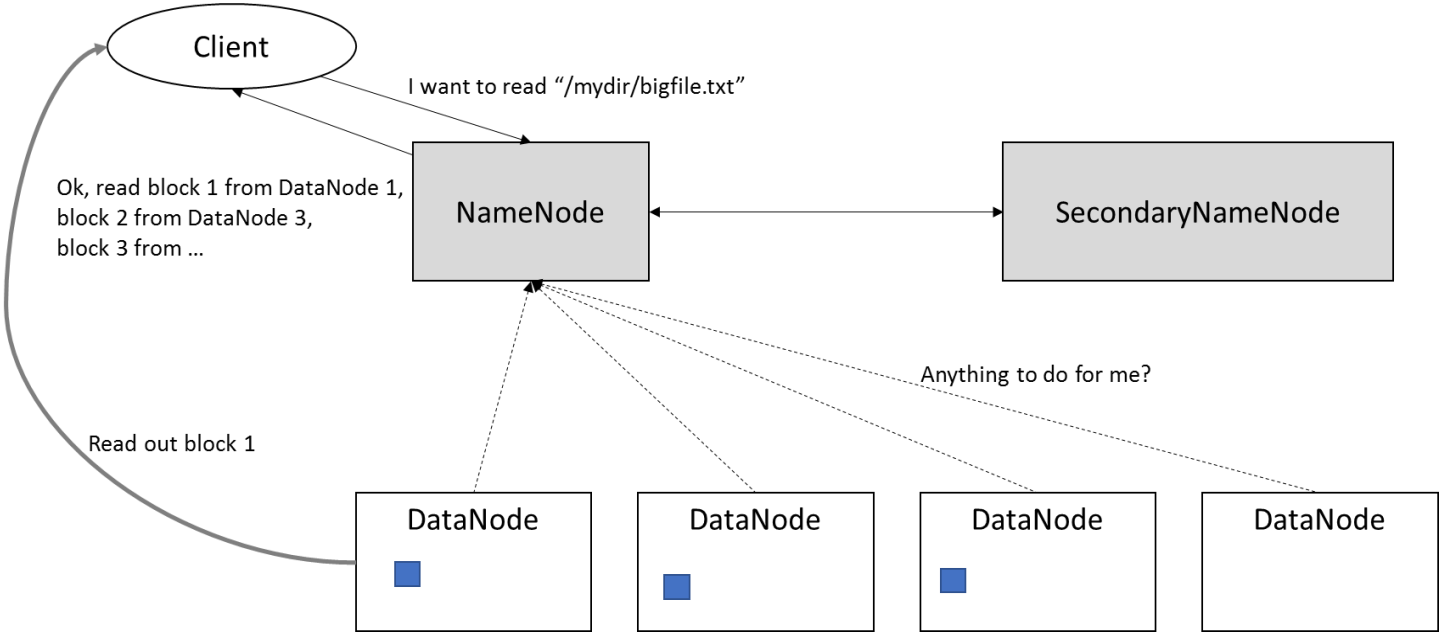
Hadoop Distributed File System (HDFS)

- NameNode
  - A server that holds all the metadata regarding the stored files (i.e., registry)
  - Manages incoming file system operations
  - Maps data blocks (parts of files) to DataNodes
- DataNode
  - Handles file read and write requests
  - Creates, deletes, and replicates data blocks among their disk drives
  - Continuously loop, asking the NameNode for instructions
- Note: size of one data block is typically 64 megabytes

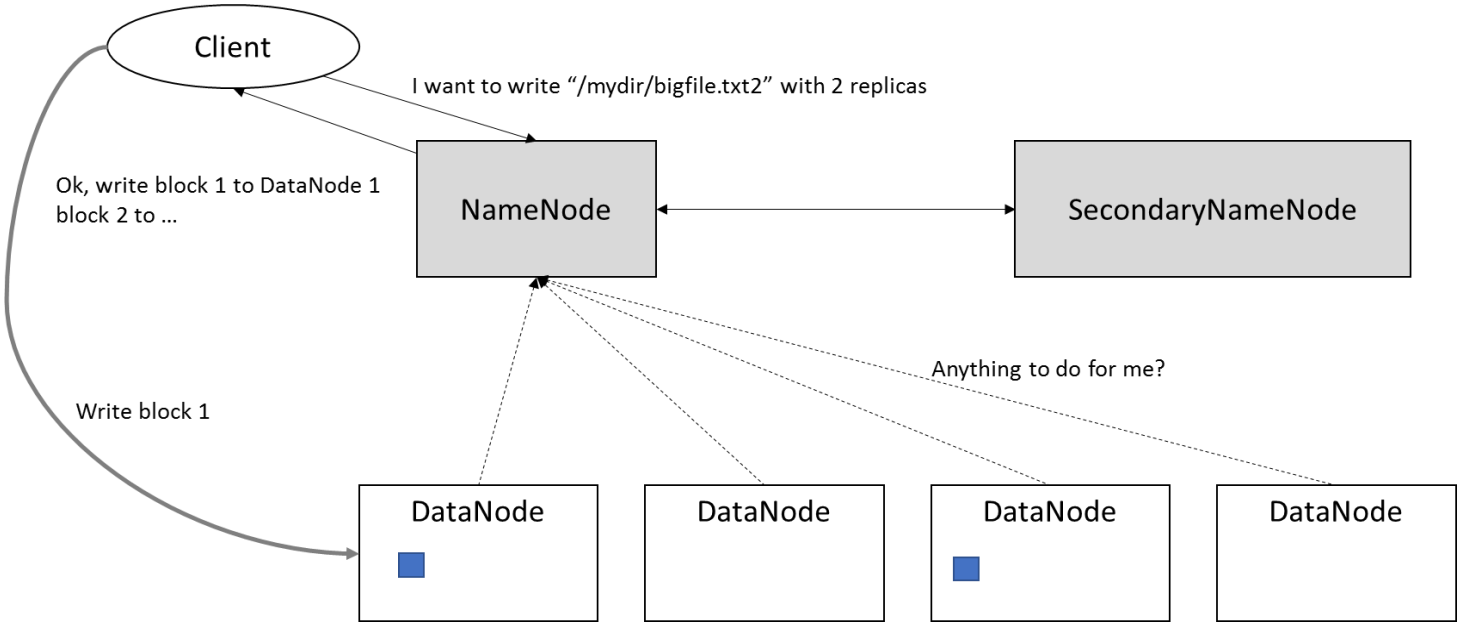
Hadoop Distributed File System (HDFS)



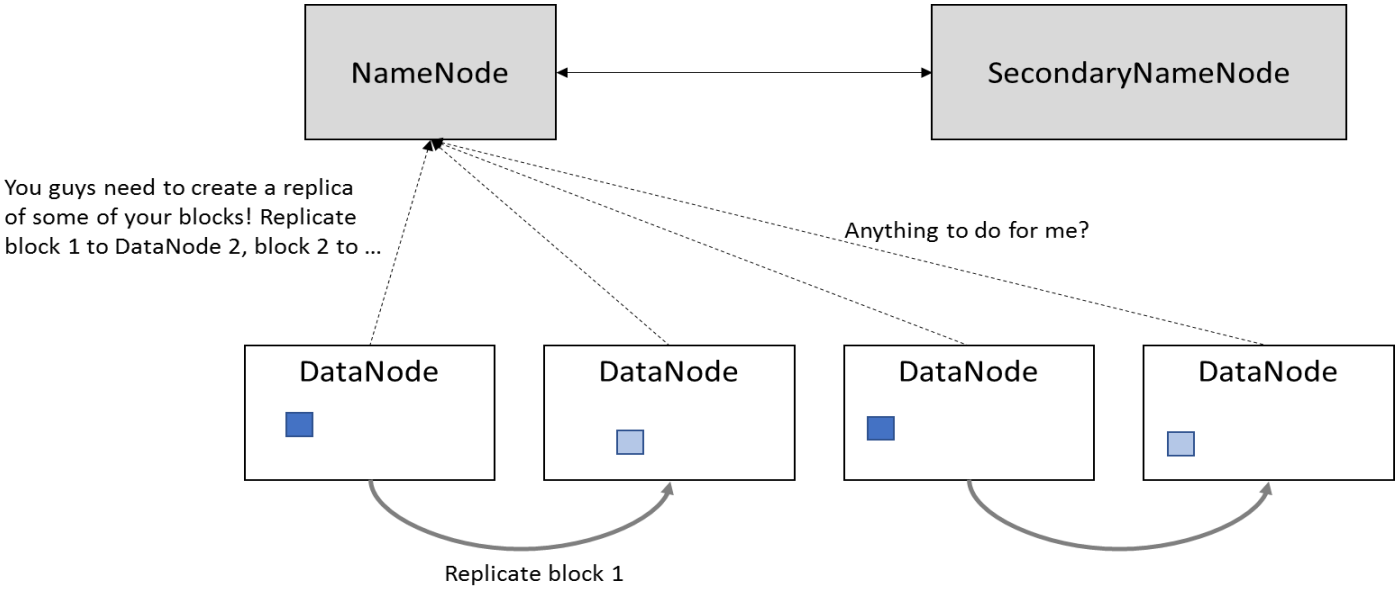
Hadoop Distributed File System (HDFS)



Hadoop Distributed File System (HDFS)



Hadoop Distributed File System (HDFS)



Hadoop Distributed File System (HDFS)

HDFS provides a native Java API to allow for writing Java programs that can interface with HDFS

```
String filePath = "/data/all_my_customers.csv";
Configuration config = new Configuration();
org.apache.hadoop.fs.FileSystem hdfs =
org.apache.hadoop.fs.FileSystem.get(config);
org.apache.hadoop.fs.Path path = new
org.apache.hadoop.fs.Path(filePath);
org.apache.hadoop.fs.FSDataInputStream inputStream =
hdfs.open(path);
byte[] received = new byte[inputStream.available()];
inputStream.readFully(received);
```

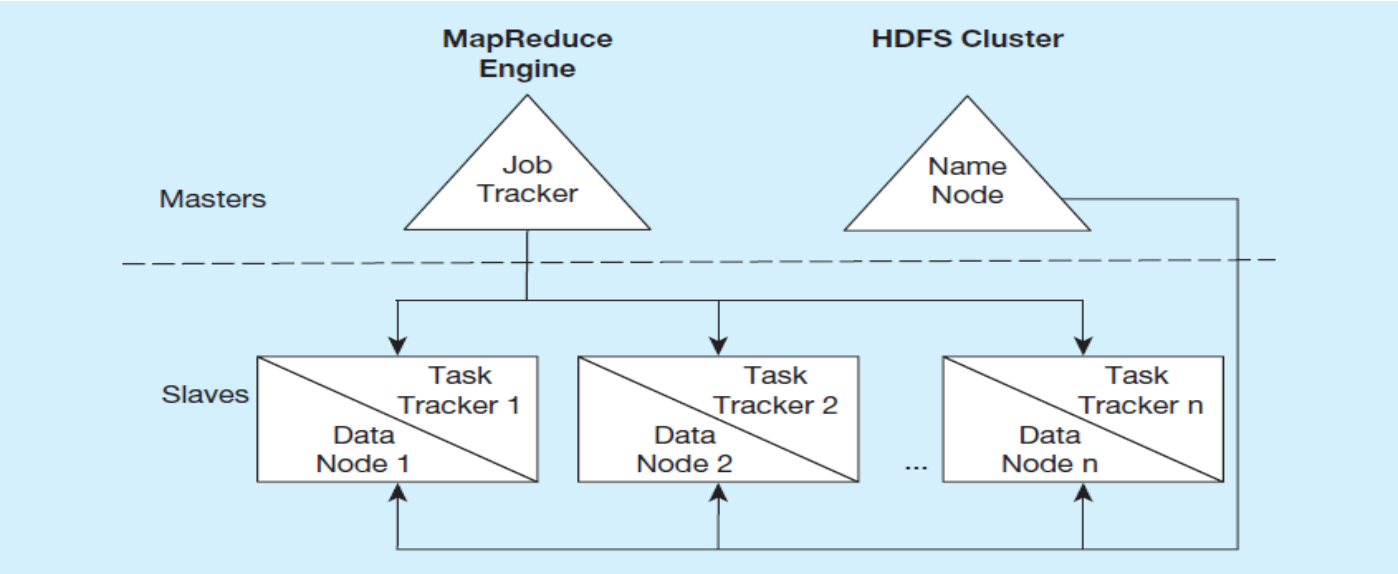
Hadoop Distributed File System (HDFS)

```
// ...
org.apache.hadoop.fs.FSDataInputStream inputStream =
hdfs.open(path);
byte[] buffer=new byte[1024]; // Only handle 1KB at once
int bytesRead;
while ((bytesRead = in.read(buffer)) > 0) {
    // Do something with the buffered block here
}
```

Hadoop Distributed File System (HDFS)

`hadoop fs -mkdir mydir`	Create a directory on HDFS
`hadoop fs -ls`	List files and directories on HDFS
`hadoop fs -cat myfile`	View a file's content
`hadoop fs -du`	Check disk space usage on HDFS
`hadoop fs -expunge`	Empty trash on HDFS
`hadoop fs -chgrp mygroup myfile`	Change group membership of a file on HDFS
`hadoop fs -chown myuser myfile`	Change file ownership of a file on HDFS
`hadoop fs -rm myfile`	Delete a file on HDFS
`hadoop fs -touchz myfile`	Create an empty file on HDFS
`hadoop fs -stat myfile`	Check the status of a file (file size, owner, etc.)
`hadoop fs -test -e myfile`	Check if a file exists on HDFS
`hadoop fs -test -z myfile`	Check if a file is empty on HDFS
`hadoop fs -test -d myfile`	Check if myfile is a directory on HDFS

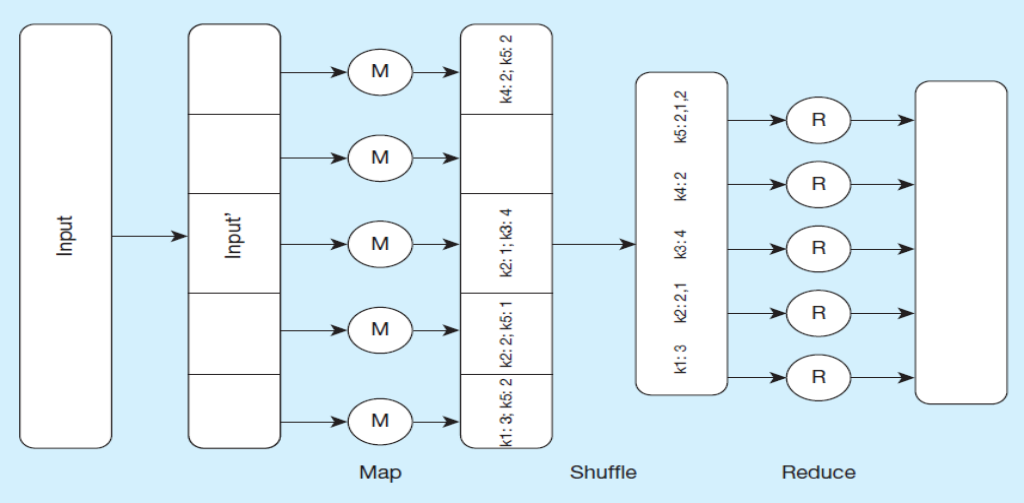
MapReduce



MapReduce

- Enables parallelization of data storage and computational problem solving via many commodity servers
- Programmers don't have to be experts at parallel processing
- Core idea – divide a computing task so that multiple nodes can work on it at the same time
- Each node works on local data doing local processing
- Two stages:
  - Map stage – divide for local processing
  - Reduce stage – integrate the results of the individual map processes

Schematic Representation of MapReduce



MapReduce: Simplified Data Processing on Large Clusters, Jeff Dean, Sanjay Ghemawat, Google, I n c., https://research.google.com/archive/mapreduce-osdi04-slides/index-auto-0007.html. Courtesy of the authors.

Other Hadoop Components

- Pig
  - Created by Yahoo developers
  - A tool that integrates a scripting language and an execution environment intended to simplify the use of MapReduce
  - Useful development tool
- Hive
  - Created by the Data Infrastructure Team at Facebook
  - Supports management and querying of large data sets and simplifies the use of Mapreduce
  - HiveQL – SQL-like language for managing Hadoop data
  - Useful for ETL tasks
- HBase
  - A wide-column store database that runs on top of HDFS
  - Not as popular as Cassandra

Sample CSV Dataset

```
111,M,150000,47401,40
123,M,10000,47408,25
456,M,100000,47405,35
222,F,125000,47401,50
345,F,20000,47408,35
567,F,250000,47403,40
678,M,175000,47403,25
789,M,300000,47405,32
333,M,30000,47408,38
444,M,75000,47401,28
```

Sample data in CSV (comma separated values). This is just a text file.

Pig Script

```
data1 = LOAD '/user/Group1/MDMSample.csv' USING PigStorage (',')
AS (userid: int, gender:chararray, salary:int, zip:chararray, age:int);


DUMP data1;
```

LOAD reads the data as tuples. USING specifies the separator between fields. AS specifies the names and data types for each item on a line. DUMP returns the results.

Sample Pig Script 2

```
data1 = LOAD '/user/Group1/MDMSample.csv' USING PigStorage (',')
AS (userid: int, gender:chararray, salary:int, zip:chararray, age:int);

filtered_data1 = FILTER data1 by (zip == '47401' or zip == '47408');
projected_filtered_data1 = FOREACH filtered_data1 GENERATE userid, salary, age;

DUMP projected_filtered_data1;
```

FILTER is like a WHERE clause in SQL. FOREACH loops through the rows. GENERATE specifies the values returned (kind of like the SELECT clause in SQL).

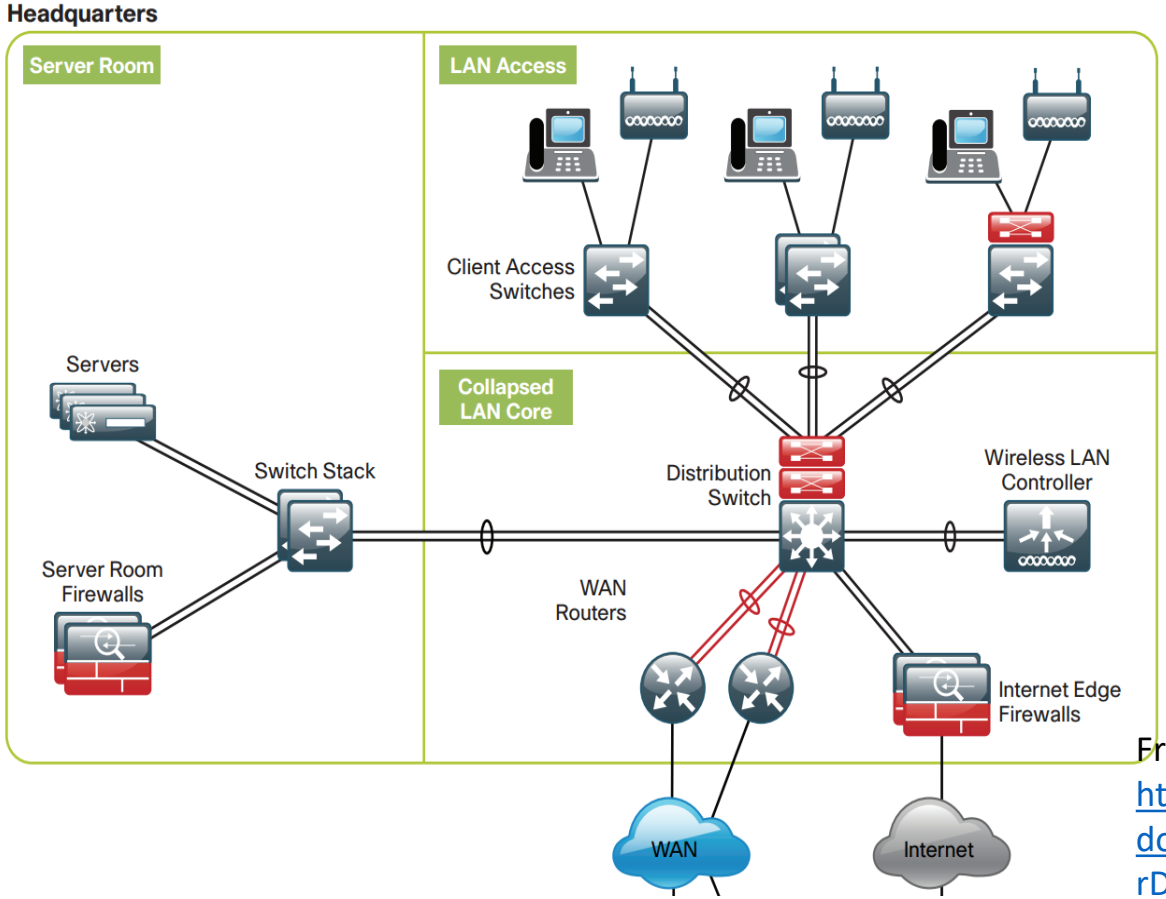
Sample Hive Script

```
CREATE TABLE customer
(userid INT, gender STRING, salary INT, zip STRING, age INT) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

Hive has a SQL-like language. Like Pig, it can apply to CSV text files and incorporated Schema on Read.

On-Premise Datacenter

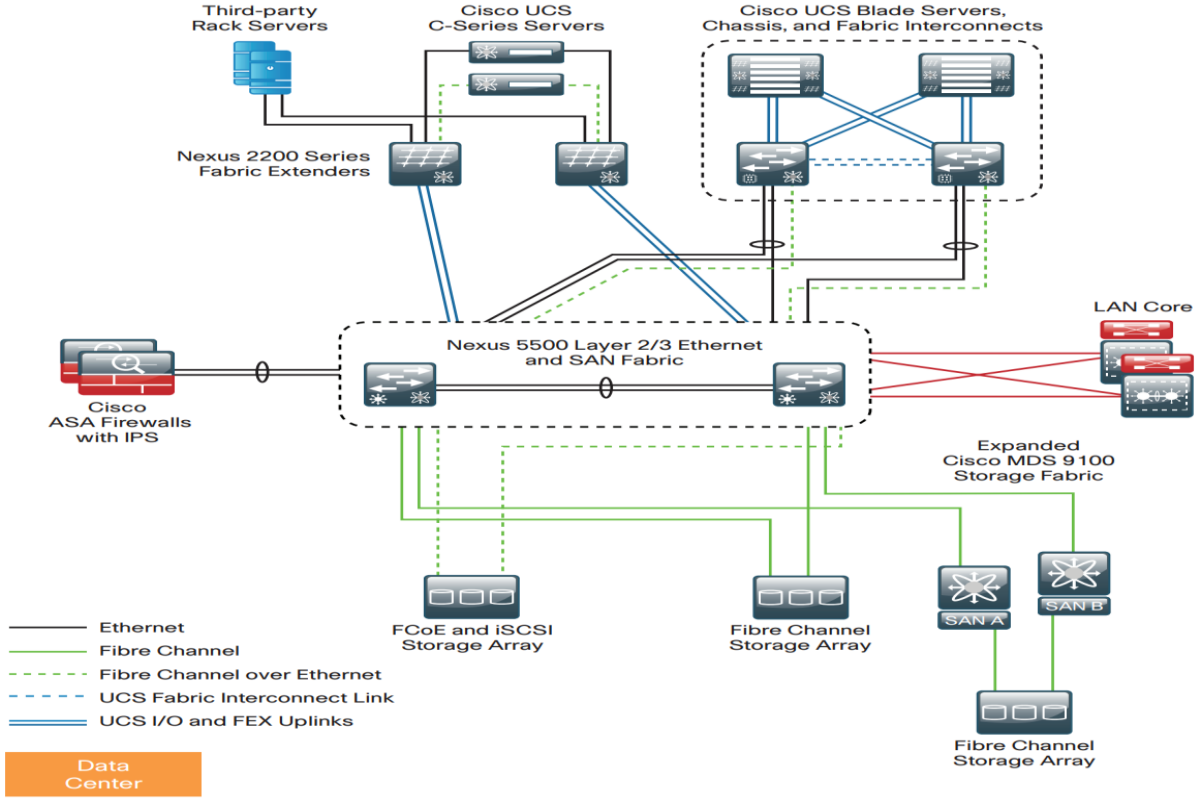
https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCInfra_1.html



On Premise Data Center

From: https://www.cisco.com/c/dam/en/us/td/docs/solutions/CVD/Aug2014/DataCenterDesignSummary-AUG14.pdf

On Premise Data Center



From: https://www.cisco.com/c/dam/en/us/td/docs/solutions/CVD/Aug2014/DataCenterDesignSummary-AUG14.pdf

Cloud based Management Services

Cloud computing

Provisioning/acquiring computing services on demand using centralized resources accessed through public Internet or private networks

Infrastructure-as-a-Service (IaaS)

Cloud service involving hardware and various types of systems software resources

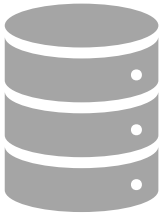
Platform-as-a-Service (PaaS)

Cloud service involving hardware and various types of systems software resources

Cloud based Management Services

Software-as-a-Service (SaaS)

Cloud service involving software solutions/applications such as Microsoft Word or Zoom



Database-as-a-Service (DBaaS)

Cloud service involving data management service such as Azure SQL

Benefits of Cloud Based Services

- No need for initial investments in hardware, physical facilities, and systems software
- Significantly lower need for internal expertise in the management of the database infrastructure
- Better visibility of overall costs of data management
- Increased level of flexibility (elasticity) in situations when capacity needs to fluctuate significantly
- Allows organizations to explore new data management technologies more easily
- Mature cloud service providers have expertise to provide a high level of availability, reliability, and security

Downside of Cloud Based Services

- Existing systems do not yet provide capacity using a model that would automatically adapt to the changing requirements targeting the system
- Current systems are not yet providing full consistency guarantees in a highly distributed environment
- Live migration is still a challenging task that requires manual planning, initiation, and management
- It is challenging to be able to monitor the extent to which cloud providers are maintaining their Service Level Agreement (SLA) commitments
- DBaaS solutions are still struggling to find fully scalable models for providing ACID support for transactions
- Continuous monthly costs

Microsoft Azure Services

- Directory: https://azure.microsoft.com/en-us/services/
- Compute: https://azure.microsoft.com/en-us/services/#compute
- Databases: https://azure.microsoft.com/en-us/services/#databases
- Storage: https://azure.microsoft.com/en-us/services/storage/

Microsoft Azure Database Services

- Azure SQL
- Azure CosmosDB – turnkey globally distributed database
- Azure SQL Database Edge – SQL engine on a small form factor computer with ARM or x64 processor
- Table Storage- NoSQL key value storage
- Azure Cache for Redis- in memory data store
- Other databases as a service: PostgreSQL, MySQL, MariaDB
- Can also install other databases on an Azure virtual machine

Azure Compute

Compute

Access cloud compute capacity and scale on demand—and only pay for the resources you use

Learn more >

Virtual Machines
Provision Windows and Linux virtual machines in seconds

Service Fabric
Develop microservices and orchestrate containers on Windows or Linux

Container Instances
Easily run containers on Azure without managing servers

SQL Server on Virtual Machines
Host enterprise SQL Server apps in the cloud

SAP HANA on Azure Large Instances
Run the largest SAP HANA workloads of any hyperscale cloud provider

Virtual Machine Scale Sets
Manage and scale up to thousands of Linux and Windows virtual machines

Mobile Apps
Build and host the backend for any mobile app

Linux Virtual Machines
Provision virtual machines for Ubuntu, Red Hat, and more

Azure CycleCloud
Create, manage, operate, and optimize HPC and big compute clusters of any scale

Azure Kubernetes Service (AKS)
Simplify the deployment, management, and operations of Kubernetes

App Service
Quickly create powerful cloud apps for web and mobile

Batch
Cloud-scale job scheduling and compute management

Cloud Services
Create highly-available, infinitely-scalable cloud applications and APIs

Azure Functions
Process events with serverless code

Web Apps
Quickly create and deploy mission critical web apps at scale

API Apps
Easily build and consume Cloud APIs

Windows Virtual Desktop
The best virtual desktop experience, delivered on Azure

Azure VMware Solution by CloudSimple
Run your VMware workloads natively on Azure

Azure Database

Databases

Learn more >

Support rapid growth and innovate faster with secure, enterprise-grade, and fully managed database services

Azure API for FHIR

Easily create and deploy a FHIR service for health data solutions and interoperability

Azure SQL Database

Managed, intelligent SQL in the cloud

Azure Cache for Redis

Power applications with high-throughput, low-latency data access

Azure Database for PostgreSQL

Managed PostgreSQL database service for app developers

Azure Database for MySQL

Managed MySQL database service for app developers

Azure SQL Database Edge PREVIEW

Small-footprint, edge-optimized data engine with built-in AI

SQL Server on Virtual Machines

Host enterprise SQL Server apps in the cloud

Azure Cosmos DB

Globally distributed, multi-model database for any scale

Table Storage

NoSQL key-value store using semi-structured datasets

Azure Database for MariaDB

Managed MariaDB database service for app developers

Azure Database Migration Service

Simplify on-premises database migration to the cloud

Azure Storage

Storage

Get secure, massively scalable cloud storage for your data, apps, and workloads

Learn more >

Storage Accounts
Durable, highly available, and massively scalable cloud storage

StorSimple
Lower costs with an enterprise hybrid cloud storage solution

Blob Storage
REST-based object storage for unstructured data

Managed Disks
Persistent, secured disk storage for Azure virtual machines

File Storage
File shares that use the standard SMB 3.0 protocol

Avere vFXT for Azure
Run high-performance, file-based workloads in the cloud

Azure HPC Cache
File caching for high-performance computing (HPC)

Storage Explorer
View and interact with Azure Storage resources

Azure Data Share
A simple and safe service for sharing big data with external organizations

Azure Backup
Simplify data protection and protect against ransomware

Azure Data Lake Storage
Massively scalable, secure data lake functionality built on Azure Blob Storage

Disk Storage
Persistent, secured disk options supporting virtual machines

Queue Storage
Effectively scale apps according to traffic

Data Box
Appliances and solutions for data transfer to Azure and edge compute

Azure FXT Edge Filer
Hybrid storage optimization solution for HPC environments

Archive Storage
Industry leading price point for storing rarely accessed data

Azure NetApp Files
Enterprise-grade Azure file shares, powered by NetApp

Amazon Web Services

- https://aws.amazon.com/
- Free tier: https://aws.amazon.com/free/?nc2=h_ql_pr_ft&all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc
- EC2 – computing instances https://aws.amazon.com/ec2/instance-types/
- Pricing estimate: https://calculator.aws/#/addService

AWS Compute

Amazon EC2
Virtual servers in the cloud

Amazon EC2 Auto Scaling
Scale compute capacity to meet demand

Amazon Lightsail
Launch and manage virtual private servers

AWS Batch
Run batch jobs at any scale

AWS Elastic Beanstalk
Run and manage web apps

AWS Lambda
Run code without thinking about servers

AWS Outposts
Run AWS infrastructure on-premises

AWS Serverless Application Repository
Discover, deploy, and publish serverless applications

AWS Wavelength
Deliver ultra-low latency applications for 5G devices

VMware Cloud on AWS
Build a hybrid cloud without custom hardware

AWS Databases

Amazon Aurora High performance managed relational database	**Amazon DynamoDB** Managed NoSQL database	**Amazon DocumentDB (with MongoDB compatibility)** Fully managed document database
Amazon ElastiCache In-memory caching system	**Amazon Managed Apache Cassandra Service** Managed Cassandra-compatible database	**Amazon Neptune** Fully managed graph database service
Amazon Quantum Ledger Database (QLDB) Fully managed ledger database	**Amazon RDS** Managed relational database service for MySQL, PostgreSQL, Oracle, SQL Server, and MariaDB	**Amazon RDS on VMware** Automate on-premises database management
Amazon Redshift Fast, simple, cost-effective data warehousing	**Amazon Timestream** Fully managed time series database	**AWS Database Migration Service** Migrate databases with minimal downtime

AWS Storage

Amazon Simple Storage Service (S3)
Scalable storage in the cloud

Amazon Elastic Block Store (EBS)
EC2 block storage volumes

Amazon Elastic File System (EFS)
Fully managed file system for EC2

Amazon FSx for Lustre
High-performance file system integrated with S3

Amazon FSx for Windows File Server
Fully managed Windows native file system

Amazon S3 Glacier
Low-cost archive storage in the cloud

AWS Backup
Centralized backup across AWS services

AWS Snow Family
Physical devices to migrate data into and out of AWS

AWS Storage Gateway
Hybrid storage integration

CloudEndure Disaster Recovery
Highly automated disaster recovery

Google Cloud Services

- https://cloud.google.com/
- Compute: https://cloud.google.com/products/compute
- Databases: https://cloud.google.com/products/databases
- Storage: https://cloud.google.com/products/storage

Google Databases

DATABASE TYPE	COMMON USES	GCP PRODUCT
Relational	Compatibility Transactions Complex queries Joins	Cloud SQL Cloud Spanner
NoSQL / Nonrelational	Time series Streaming Mobile Web IoT Offline sync Caching Low latency	Cloud Bigtable Cloud Firestore Firebase Realtime Database Cloud Memorystore

Google Compute

Offering	Common uses	Industry
Compute Engine Scalable, high-performance and general purpose VMs.	• LOB apps • Web hosting • Enterprise apps • Databases • Most workloads	• Education • Energy • Financial services • Gaming • Government • Healthcare • Life sciences • Media and entertainment • Retail • Telecommunications
Migrate for Compute Engine Server and VM migration to Compute Engine (formerly Velostrata).	Migrate applications from on-premises, multiple data centers, or clouds to Google Cloud.	
Cloud GPUs GPUs for machine learning, scientific computing, and 3D visualization.	• Machine learning • Medical analysis • Seismic exploration • Video transcoding • Graphic visualization • Scientific simulations	• Gaming • Information technology • Life sciences • Media and entertainment
Preemptible VMs Affordable, short-lived compute instances suitable for batch jobs and fault-tolerant workloads.	• Short-lived or fault-tolerant workloads • Financial modeling • Rendering • Media transcoding • Manufacturing design • Hadoop and big data • Continuous integration • Web crawling	• Energy • Finance • Healthcare • Media and entertainment • Pharmaceuticals
Shielded VMs Hardened virtual machines	• Defend against rootkits and bootkits • Protect enterprise workloads	• Financial services • Logistics

Google Storage

Products	Good for
OBJECT OR BLOB STORAGE	
Cloud Storage Reliable object storage with global edge-caching and instant data access.	• Stream videos • Image and web asset libraries • Data lakes
BLOCK STORAGE	
Persistent disk High-performance block storage for virtual machines and containers.	• Disks for virtual machines • Sharing read-only data across multiple virtual machines • Rapid, durable backups of running virtual machines • Storage for databases
Local SSD Ephemeral locally-attached block storage for virtual machines and containers.	• Flash-optimized databases • Hot caching layer for analytics • Application scratch disk
ARCHIVAL STORAGE	
Cloud Storage Ultra low-cost archival storage with online access speeds.	• Backups • Media archives • Long-tail content • Meet regulation or compliance requirements
FILE STORAGE	
Cloud Filestore Fully managed, scalable file storage with predictable performance.	• Home directories • Rendering and media processing • Application migrations
MOBILE APPLICATION	
Cloud Storage for Firebase	• User-generated content

In Class Labs

- QwikLab Google Cloud Datastore
  - Go to link and join the class. This lab is free.
  - Please upload a print screen of the completion page or the email confirmation of completion in Blackboard

Data 604 Data Management

DATA 604 Data Management (03.12333) SP2021

Week 10 In Class Lab 4/5/2021 Google Cloud Storage

Assignment Content

QwikLab Google Cloud Storage

- Go to link , create an account if needed and join the class. This lab is free.
–Please upload a print screen of the completion page or the email confirmation of completion here in Blackboard

(1) Click this link to access the lab

Details & Information

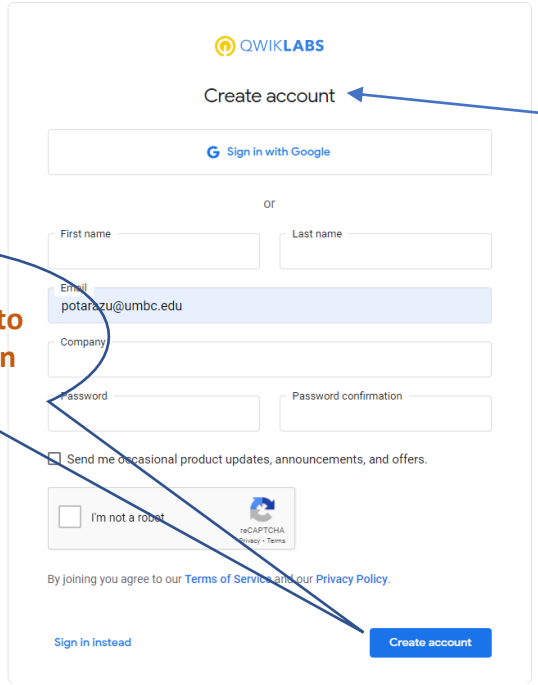
Assessment due date
4/5/21, 11:59 PM (EDT)

Attempts
Unlimited

ts 20 points

Use this space to build your submission.
You can add text, images, and files.

Add Content



(2a) Alternately you can use any gmail id you may have and sign in with Google

(2) Enter you Student email id to setup and click on Create account

(3) You will receive an email to confirm the account creation. Click the link and confirm you account creation. Note: If you miss this step, you may not be able to sign in.

Cloud vs On-Premise – Things to consider

- Existing Resources (facilities, staff, contract agreements, hardware)
- Start up or mature brick and mortar operation?
- Growth and long term expectations
- Monthly, annual, and five year expected costs
- Company goals and competencies

Estimating Cloud Costs

- Calculate initial, monthly, yearly and five year costs
  - Services, storage, software, CPU, and memory
  - Staff to configure and maintain cloud services
- Amazon Pricing Calculator: https://calculator.aws/#/addService
- Microsoft Azure Pricing Calculator: https://azure.microsoft.com/en-us/pricing/calculator/
- Google Cloud Pricing: https://cloud.google.com/products/calculator/
- IBM: https://www.ibm.com/cloud/pricing

Estimating On-Premise Costs

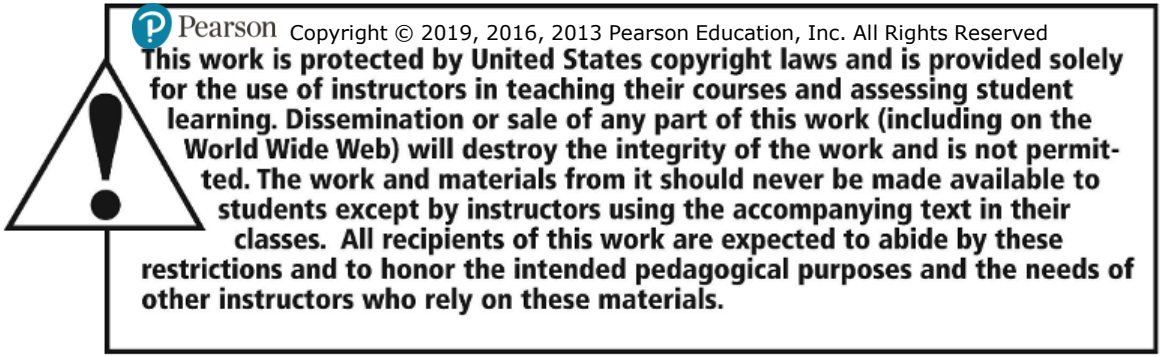
- Companies may have negotiated contracts in place, check existing contracts first.
- For no contracts, conduct market research for estimates:
  - Government - GSA https://www.gsaadvantage.gov
  - Amazon or Newegg for HW/SW estimates

On Premise Pricing

- Costs include:
  - Hardware (server, storage, racks, accessories)
  - Maintenance
  - Software licenses
  - Staff to install and maintain
  - Facilities (rent, cooling, power, security, backup)
  - Disaster Recovery (backup tape, storage, failover center)
  - Training
- Typically calculate initial, yearly and five year costs

Homework 3 Assigned

- Pricing Estimate On Premise vs Cloud
- Assignment is posted in Blackboard
- Individual Assignment
- Due 4/13/2021 through Blackboard


Pearson
Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved
This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.