Review For Midterm Exam #2

Midterm →

Covers material since the first midterm exam in March

How can I best prepare?

- View the recorded lectures if you have not yet viewed
- Review slides from each lecture
- Review your homework (# 5, 6, 7)
- Review Weekly Quizzes

NOTE →

- Take the exam at 3:00 p.m. on Wednesday, April 29.
  (Unless authorized for an alternate time...)
- There is NO LECTURE on Friday, May 1 ("reading" day)

Review For Midterm Exam # 2

- Approximately 20 questions
- 50 minutes*
- Exam administered via Moodle
- 8-10 questions per page (to reduce calls to Moodle)
- Multiple choice, fill-in, matching, True/False
- Open book, open notes, open internet
- You are being trusted to do your own work. No collaboration allowed.
- The exam will open at 3:00 p.m. on Wednesday, April 29
  - (Unless you are authorized to use an alternate copy)
- Once you start the exam, it will close automatically when the timer is up
- Correct answers will not be available until the exam closes

* Certain students are allowed accommodations for extended time. If this is you, and if you have emailed me in the past two weeks requesting this, I will send you an email with instructions and the password to open your copy of the exam.

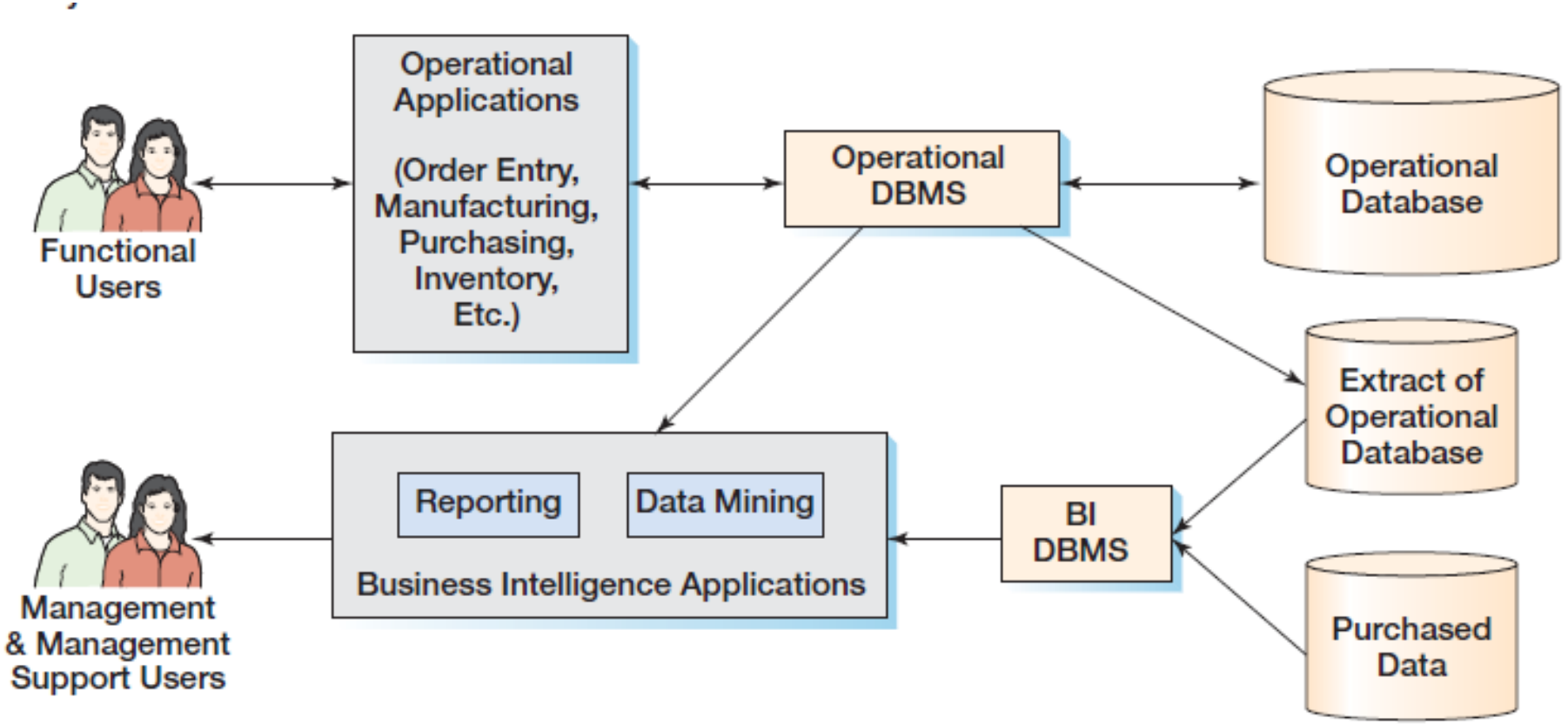
Review For Midterm Exam # 2

- Business Intelligence
- Big Data Concepts
- NoSQL Concepts
  - 4 models of NoSQL databases
  - MongoDB
  - Cassandra
- Hadoop Concepts

Review For Midterm Exam # 2

- Business Intelligence
- How are BI activities different from Operational/Transactional Activities
  - In running a business?
  - In using a database?
  - OLTP versus DSS
- Who in an organization performs BI activities? For what purpose?

Review For Midterm Exam # 2

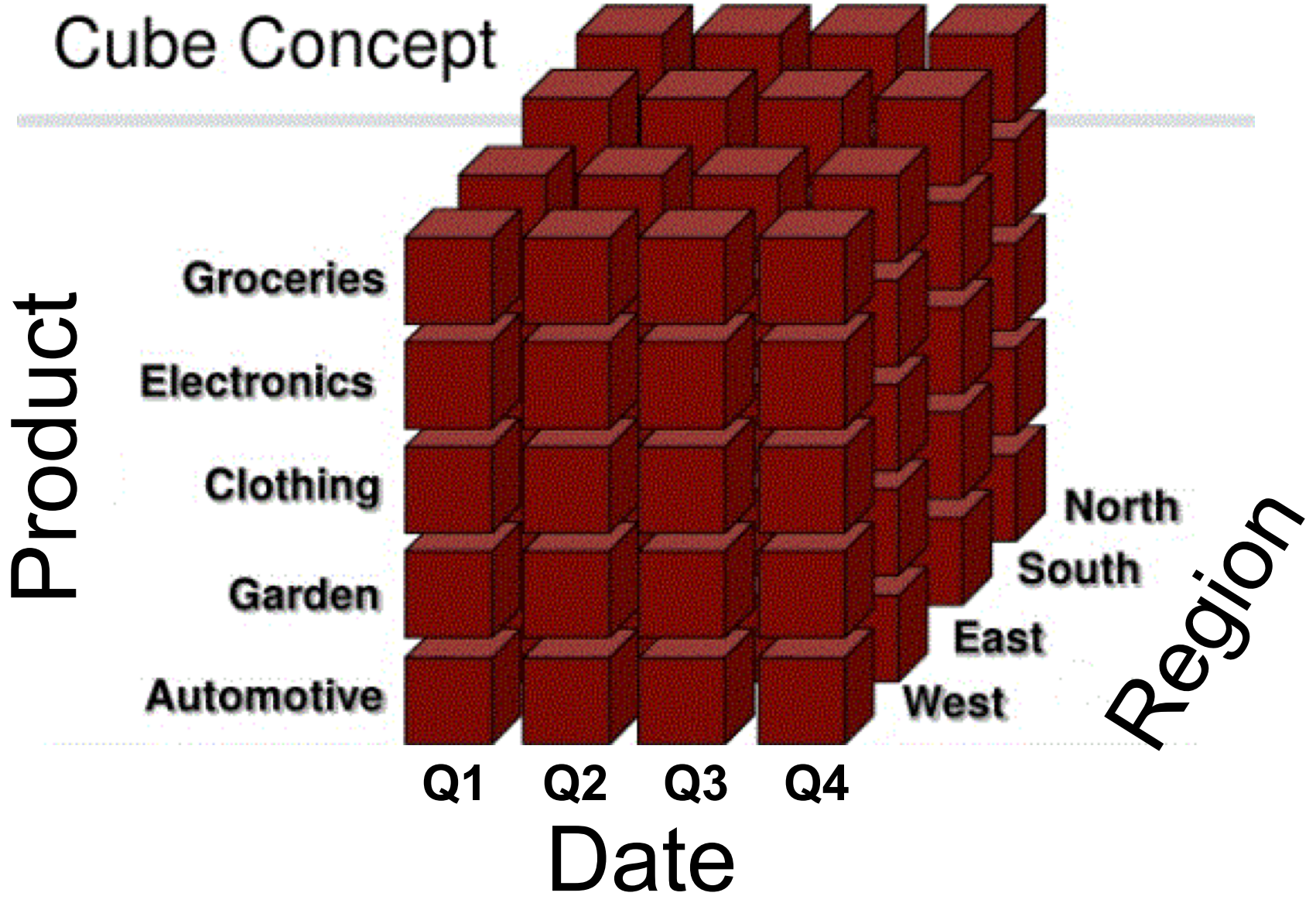


Review For Midterm Exam # 2

- Two main types of BI systems
  - Reporting
  - Data Mining

Review For Midterm Exam # 2

- OLAP – BI Analytics
  - What does the term OLAP mean?
  - What is a CUBE and how is it used in BI analytics?
  - What are Roll-up and Drill-down in terms of cube processing?


Cube Concept
Product
Groceries
Electronics
Clothing
Garden
Automotive
North
South
East
West
Region
Q1
Q2
Q3
Q4
Date

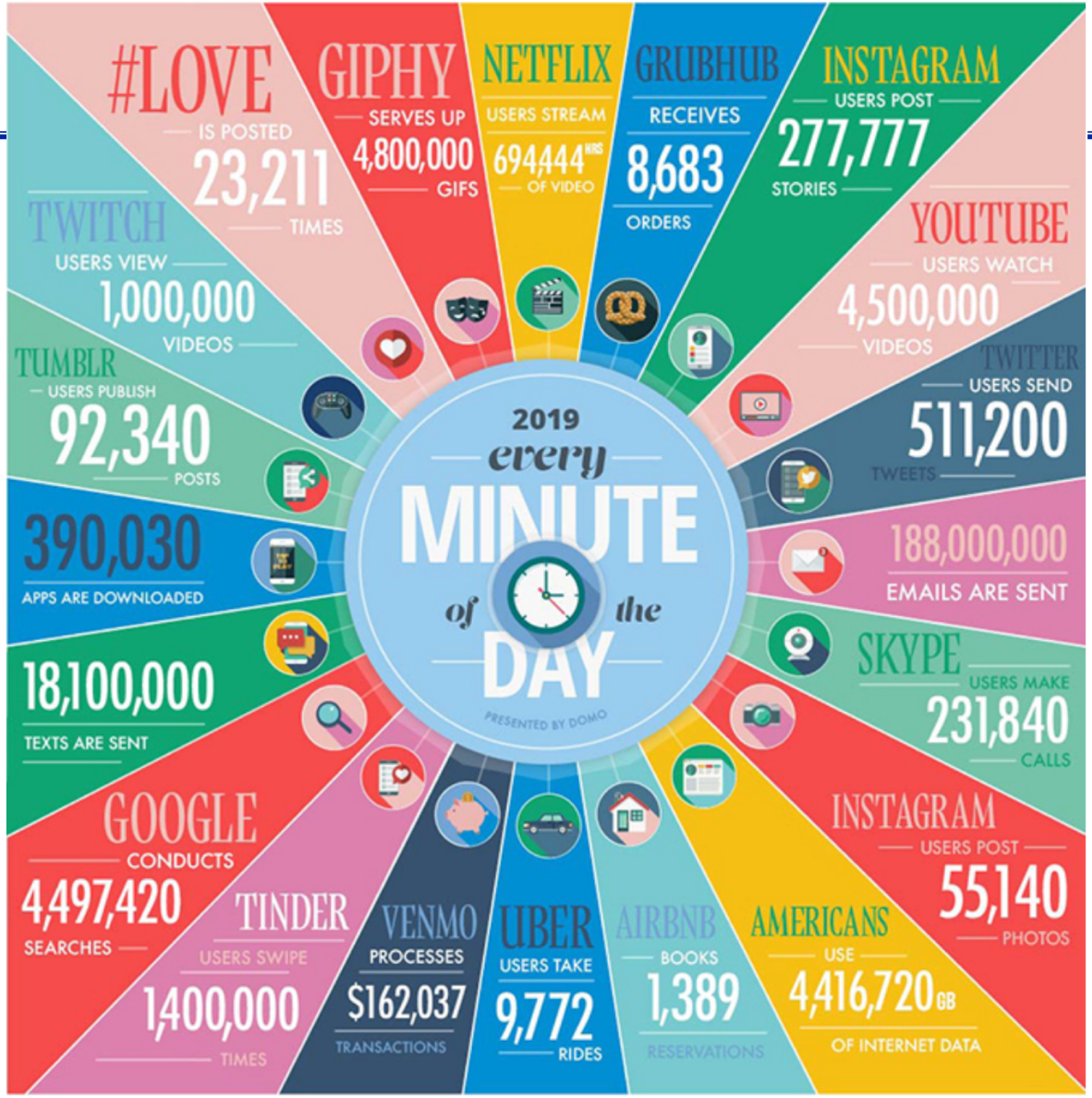
Why use a CUBE?

- Aggregates are calculated during cube refresh
- Refresh is typically scheduled to run daily following the daily warehouse ETL process
- Avoids calculating the same sums over and over
- Pre-calculated sums are based on customer usage of most commonly combined dimensions
- Cubes store summed facts at a higher grain than the base, atomic fact tables
- THEREFORE: queries against the cube(s) run much faster than queries against the fact tables

Review For Midterm Exam # 2

- Big Data
  - What does the term “big data” really mean?
  - What are the three V’s and what do they have to do with Big Data?
  - What is driving the Big Data explosion?
    - Smart Phones
    - Apps
    - The Internet of Things
    - Digital Commerce
    - Online Entertainment
    - Cloud Computing
    - Social Media

Review


2019
every
MINUTE
of the
DAY
PRESENTED BY DOMO
#LOVE
IS POSTED
23,211
TIMES
GIPHY
SERVES UP
4,800,000
GIFS
NETFLIX
USERS STREAM
694,444 HRS
OF VIDEO
GRUBHUB
RECEIVES
8,683
ORDERS
INSTAGRAM
USERS POST
277,777
STORIES
TWITCH
USERS VIEW
1,000,000
VIDEOS
YOUTUBE
USERS WATCH
4,500,000
VIDEOS
TUMBLR
USERS PUBLISH
92,340
POSTS
TWITTER
USERS SEND
511,200
TWEETS
390,030
APPS ARE DOWNLOADED
188,000,000
EMAILS ARE SENT
18,100,000
TEXTS ARE SENT
SKYPE
USERS MAKE
231,840
CALLS
GOOGLE
CONDUCTS
4,497,420
SEARCHES
INSTAGRAM
USERS POST
55,140
PHOTOS
TINDER
USERS SWIPE
1,400,000
TIMES
VENMO
PROCESSES
$162,037
TRANSACTIONS
UBER
USERS TAKE
9,772
RIDES
AIRBNB
BOOKS
1,389
RESERVATIONS
AMERICANS
USE
4,416,720 GB
OF INTERNET DATA

Review For Midterm Exam # 2

- Big Data
  - Terabytes – Petabytes – Exabytes – Zettabytes
- **The Relational Problem**
  - The Relational Database: 40+ year-old technology
    - Tables, Rows, Columns, Keys, Joins, Commits
    - ACID Transaction Compliance
    - RDBMS Software Choices
  - Traditional Database Server Architecture
    - Memory, CPU, Storage
    - The cost of scaling “UP”

Challenges

- How can we collect, process and store so much data?
- How can we wade through so much data in a timely manner?
- How can we analyze so much data and gain meaningful insights?
- Can my existing systems/architectures handle this?
  - Why not?
- Can my existing staff (skills & tools) make the transition?

Alternatives for handling Big Data

- Scale “OUT”, not “UP”
- Use cheap, commodity hardware, local disk
- Use SSDs
- Seamless Fault Tolerance -- No Single Point of Failure
- Distribute the Data: Replication & Sharding
- Distribute the Processing: Parallelization
- Redundancy: Replication
  - Master-Master
  - Master-Slave
- Relaxed ACID compliance restrictions

NoSQL versus Relational

Relational

- Schema defines rigid structure
  - Tables, Rows, Columns
- Foreign Key relationships
  - Which support joins
- Uses SQL
- Maintains ACID compliance
- Normalized: store a value only once
- Clustering is a challenge
- Main DBMS players are quite expensive

NoSQL

- Stores related values in aggregates
- Flexible structure:
  - Ranges from none to some
- No joins
- Relaxed ACID compliance
- De-normalized
- Uses alternate query language
- Easily supports clustering
- Almost all players are open source

NOSQL Data Models

- **NoSQL Data Models**
  - Document Store (using XML or JSON format)
  - Graph (using node/edge structures with properties)
  - Key-Value pairs
  - Wide Columnar store (rows with dynamic columns holding key-value pairs)

MongoDB Architecture

- **MongoDB Architecture**
  - Use of JSON, BSON
    - What are they ?
    - How/Why does MongoDB use them?
  - Mongo architecture
    - Database ➔ Collection ➔ Documents
    - Each document has a primary key generated by Mongo
  - How does Mongo provide high-availability?
    - Replication, Node Failure
    - Node role management

- **MongoDB Architecture**
  - What is sharding/partitioning?
    - Hash versus Key Value Ranges

MongoDB Query Language

- **MongoDB**
  - The MongoDB query language
  - How to view databases, collections, documents
  - Common commands
    - db.<collection>.find()
    - db.<collection>.count()
    - pretty()
    - type "it" for more
    - matching conditions ($eq, $gt, $lt, $and, $or, etc.)
    - regex in find()
  - Finding a field within a field

MongoDB Query Language

The MongoDB query language

- the aggregate() method – what is it, how does it work?
- what is a pipeline when using aggregate()
- aggregate() steps:
  - $match
  - $group
  - $unwind
  - $project
  - $sort
- matching Mongo query commands to equivalent SQL commands

Cassandra Architecture

Where did Cassandra come from?

Google Big Table + Amazon Dynamo + Facebook

Cassandra Replication/Partitioning

How is it different from MongoDB?

How does cassandra balance partition distribution?

What is the replication factor?

CAP Theorem

What is it?

Why do we talk about it when we consider NoSQL systems?

Trade-Offs

Tunable Consistency

What does this mean?

How does Cassandra implement it?

- **Hadoop**
  - What makes Hadoop so special?
  - It takes advantage of all the following NoSQL features:
    - Parallelization (for throughput)
    - Distribution of Data and Compute Tasks
    - On cheap commodity hardware (physical, not virtual)
    - Local dedicated disk
    - Fault Tolerance
    - Redundancy Factor
    - Relaxed ACID compliance
    - Chunks = Large Blocks (64MB versus 4K)
    - Linear Scalability
    - Write Once Read Many
    - Open Source

- **Hadoop**
  - Its history – where did it come from?
  - Key components of the Hadoop Ecosystem
    - HDFS
    - Name node(s)
    - Compute/Data nodes
    - YARN & Tez
    - Map/Reduce
    - Tools in the environment
      - Hive
      - Pig
      - SQOOP
      - SOLR

Cloud Databases

- Who are the "big three" in Cloud Database providers?
- What trends are currently happening among cloud providers?
- What differentiates the big three from each other?
- What is the difference between a Private Cloud and a Public Cloud?
- Why would an organization choose Private Cloud versus Public Cloud?
- Advantages and disadvantages of using a public cloud database?