Q1: What is the difference between EBS and S3?

- EBS is block storage that mounts to an EC2 instance (virtual computer)
- S3 is object storage that is a scalable hard drive. S3 can be used by many virtual computers ("giant pile of storage")
- In terms of capacity, S3 is unlimited where EBS has a cap at 16TB (imagine both as warehouses where S3 has "no roof")
- Use case for S3 might be Nasa storing all of their planet images
- EBS is better for dataset that doesn't need to scale and requires a lot of read-write operations
- EBS appears as a mounted volume to an EC2 instance whereas you need separate software to access S3

Q2: When would you consider a cloud infrastructure for your data science tasks?

There are many scenarios but some include:

- When the compute load is greater than the capacity of your local computers/servers
- When you want to take advantage of cloud products like Elastic Map Reduce from Amazon
- When you have a very large dataset that can't be reasonably stored on local computers/servers
- When you're out of testing mode and suddenly need to run scaled processing but know you'll eventually ramp down

Q3: What is the difference between spot instances and reserved instances?

- Spot instances are operate like EBay where you can bid on unused EC2 instances
- With spot instances, there may be some cost savings from the dynamic bidding versus "on-demand" which is a stable price
- Reserved instances operate like movie tickets where you book your seats in advance
- With reserved instances, you have more predictable compute loads so can anticipate what you need at what time

Q4: List the names of four software packages installed on the UCB AMI

Appears to be:

- Hadoop
- Streamparse
- iPython
- PostgreSQL (pgxl-deployment-tools)