

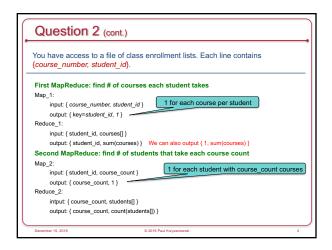
You have access to a file of class enrollment lists. Each line contains {course\_number, student\_id}.

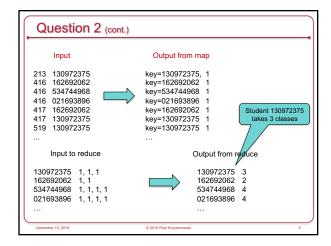
Explain how you would use MapReduce to get information on how many classes students take.

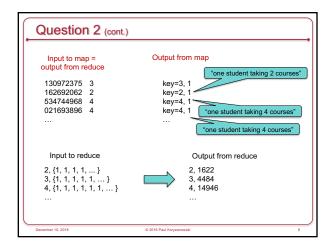
For instance, you may discover that 1,495 students are enrolled in 6 courses; 13,077 students are enrolled in 5 courses; 14,946 students are enrolled in 4 courses; and 4,484 students are enrolled in 3 courses.

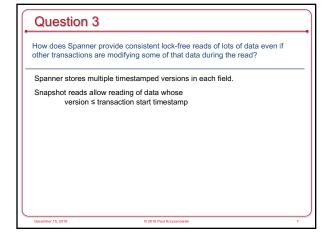
Explain each map and reduce operation. You may use pseudocode and assume that functions such as sum and count exist. Be sure to state the inputs & outputs of each step.

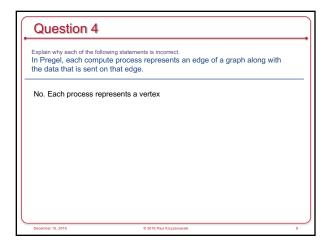
Hint: you may need more than one iteration











Cuestion 5

Explain why each of the following statements is incorrect.
A shared nothing architecture does not allow two systems to access the same NFS server.

A shared nothing architecture does not allow two systems to share the same block-level storage (e.g., cluster file system).

No credit for simply negating the statement!

Explain why each of the following statements is incorrect.
Akamai uses dynamic DNS to look up a URL and return the address of the closest caching server that has that URL in its cache.

1. Akamai uses dynamic DNS to look up domain names, not URLs! DNS does not do URL lookups

2. Akamai does not necessarily return the closest server. It will return one that is available, not loaded, likely to contain content, ... and is closest (lowest latency)

Explain why each of the following statements is incorrect.
A digital certificate contains a hash that is encrypted with the certificate owner's private key.

Encrypted hash = signature.
A digital certificate is singed by the certification authority (CA), not the owner:
The hash is encrypted with the CA's private key

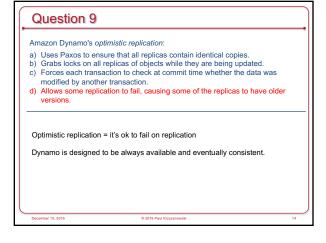
Not: encrypted with the certificate owner's public key

Explain why each of the following statements is incorrect.
A digital certificate contains a hash that is encrypted with the certificate owner's private key.

Encrypted hash = signature.
A digital certificate is singed by the certification authority (CA), not the owner:
The hash is encrypted with the CA's private key

Not: encrypted with the certificate owner's public key

#### What is an advantage of Amazon Dynamo's *virtual nodes* over simply having each physical machine be a node? a) Performance is improved because you can create a massive number of virtual nodes. b) It's cheaper because the node runs on a virtual machine that can do other things as well. c) A virtual node can be easily migrated to another system. d) You can shed load from multiple systems when you add a new physical machine.. Adding a new machine adds more virtual nodes that are scattered throughout the ring. → Each virtual node takes on some {key, value} sets from another node On average, these other nodes are spread across multiple physical systems



MapReduce can process a lot of data relatively quickly because:

a) Input data is broken up among many map workers.
b) Map workers run in parallel with reduce workers.
c) Reduce workers shrink the amount of data that needs to be processed.
d) All of the above.

MapReduce is a great example of divide-and-conquer.
(b) No. Reduce workers run after every single map worker is done
(c) Only sometimes with multiple iterations of MapReduce. In general, map workers get rid of unnecessary data and reduce workers compute results

Cuestion 11

The partitioning function in MapReduce determines:

a) How the original input data is divided into multiple shards.
b) Which reduce worker will work on a specific key.
c) How the map worker parses input data to generate a key.
d) The maximum number of values that will be assigned to each key.

The partitioning function is applied to the {key, value} results from map workers.

A column family in Bigtable is:

a) A group of columns within a single table.
b) A group of related columns among multiple tables.
c) A set of columns that are shared across multiple tables.
d) A preconfigured set of columns to make it easier to set up a Bigtable instance.

A column family is a storage unit for multiple columns in a table – usually related in function.

From a performance point, data in a row is retrieved by column families

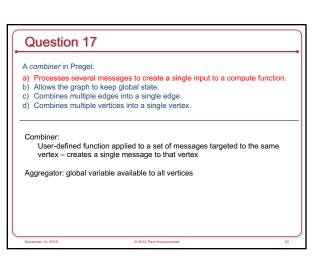
Rows in Bigtable:

a) Are always stored in a sorted order by the row key.
b) Are ordered chronologically; the last row always gets added to the end of the table.
c) Are not sorted but a row can be located by looking up its key in a hash table stored at the master.
d) Are not sorted or searchable but an application can get a list of all rows that contain data for a specific column.

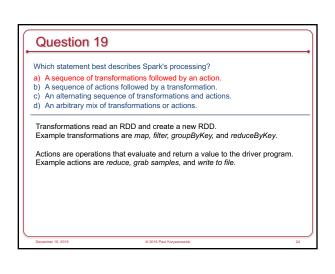
#### Question 14 Spanner transactions are kept isolated by: a) Using the TrueTime API to ensure no two transactions start at the same time. b) Using strict two-phase locking on resources. c) Using a two-phase commit protocol. d) Waiting until all previous transactions have committed. (a) Transactions may start at the same time. (c) This deals with commits, not isolation during the transaction. (d) No – but this would also deal with commits rather than isolation

#### To provide external consistency in Spanner, commits are delayed: a) Until all earlier transactions have released all their locks. b) All the locks used by the transaction are released. c) Up to the estimated clock skew between clocks in different locations. d) Until all earlier transactions have committed. External consistency = order of transactions reflects their true time order Commit wait: before a transaction commits, it acquires a commit timestamp: t = TT.now().latest t = the latest possible value of the true time across all servers in the system. Do not release any locks until that time is definitely in the past. This means waiting until the earliest possible current time on any system is greater than the transaction timestamp: TT.now().earliest > t

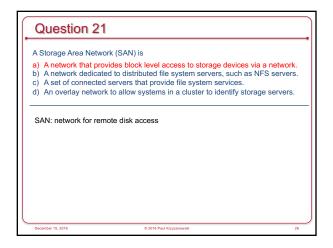
## In the Pregel framework, the entire job is complete when: a) Every compute function vote to halt. b) Every compute function entered an inactive state. c) There are no incoming messages for any compute function. d) Every compute function entered an inactive state and has no incoming messages. Voting to halt == entering an inactive state BUT: A compute function can send a message before voting to halt If a compute function receives a message, it is reawakened – placed back into an active state



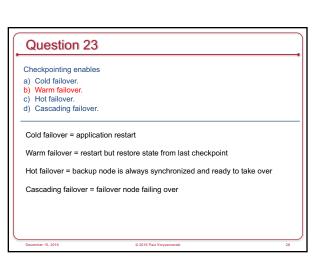
# A Spark resilient distributed dataset (RDD) is fault tolerant because: a) It is stored as replicas on multiple servers. b) It is partitioned across many servers in the cluster. c) It can be regenerated using the transformation that created it. d) It is cached in memory with a backup on the disk of the servers that created it. (a) No. An RDD is sharded but generally not replicated. (b) Yes, it could be partitioned but that does not make it fault tolerant. (d) Also possible but disk is generally used as a RAM overflow. RDDs are created by transformations and can be re-created by those same transformations.



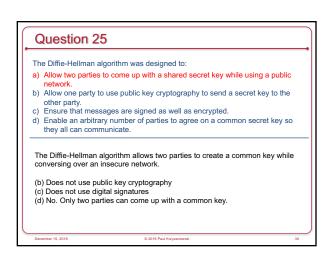
## A System Area Network (SAN) is typically: a) An overlay network that identifies the nodes that make up the cluster. b) A dedicated network that is used only for cluster management messages, such as heartbeats. c) A high bandwidth, high reliability, low latency network used to connect nodes of a cluster. d) A high-speed network used to connect processors within a multiprocessor system. High performance, low-latency, high-reliability LAN interconnect for cluster members. Examples: Infiniband, Myrinet, ethernet with Data Center Bridging



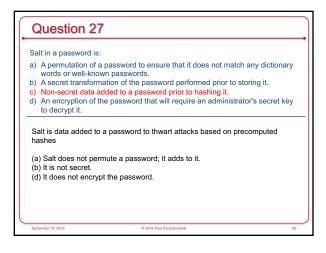
## The difference between a clustered file system and a distributed file system is a) A clustered file system is only made available to a limited set of systems. b) Distributed file systems may span multiple servers. c) Clustered file systems provide replication for fault tolerance. d) All nodes access the clustered file system at the block level. Each system maintains its own file system and uses a shared block device for storage. Distributed file system: server(s) maintain the file system and provide file system APIs for remote access (e.g., read bytes from a file, write bytes to a file, create a file, delete a file)



# Akamai's overlay network is primarily designed to: a) Enable load balancing. b) Enable load shedding. c) Provide secure transport of requests within Akamai's systems. d) Find faster routes than might be available via Internet routing. Akamai collects statistics about its servers and their availability & connectivity Core function of Akamai: content caching Akamai's dynamic DNS enables load balancing & load shedding Overlay network enables Akamai to find faster routes to origin servers



## A session key is: a) A key derived from one user's private key and another user's public key. b) A randomly generated key designed to be used for one communication session. c) A key that is used only for signing messages. d) A key that automatically expires after a certain number of messages have been sent. Session key = random, throw-away key used for one session.

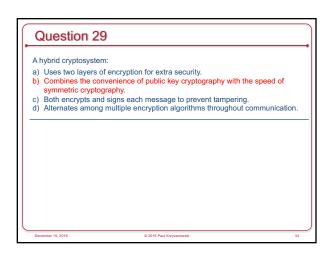


For Alice to authenticate Bob, she sends him a bunch of bytes that:

a) Bob encrypts with his private key.
b) Bob encrypts with his public key.
c) Bob encrypts with Alice's private key.
d) Bob encrypts with Alice's public key.

(a) Alice can decrypt the result using Bob's public key to validate that only Bob could have performed the encryption

(b) Anyone can encrypt the data with Bob's public key.
(c) Bob does not have access to Alice's private key.
(d) Anyone can encrypt the data with Alice's public key.



Tunneling is the technique of
a) Encrypting the data part of a packet before transmitting it.
b) Encrypting the entire packet before transmitting it.
c) Using the data part of a packet to transport another packet.
d) Signing a packet to prevent unauthorized modifications to it.

Tunneling is simply the encapsulation of one packet within the payload of another.

VPNs use tunneling and add encryption and/or signatures.

