Write a mapper/reducer function for counting words in a file.

GIS DB use case

How does MapReduce improve the reliability of distributed systems?

Give few examples of “Transformations” in RDD operations.

RDDs support two types of operations: transformations, which create a new dataset from an existing one, and actions, which return a value to the driver program after running a computation on the dataset.

Explain why query optimization is necessary

For any production database, SQL query performance becomes an issue sooner or later. Having long-running queries not only consumes system resources that makes the server and application run slowly, but also may lead to table locking and data corruption issues. So, query optimization becomes an important task.

Mention 2 advantages of HDFS’s data block approach

**Simplicity of storage management**

### Ability to store very large files

### Simple Storage mechanism for datanodes

### Fault tolerance and High Availability of HDFS

How does name-node help in recovery from failed data-nodes in the HDFS?

Name node continuously check the heartbeat signal from the data node and if it stops , then namenode consider it as dead and reaches out to another live datanode of same data,

To recover the lost data the , NameNode consults its in-memory metadata to determine which data blocks were managed by the node.

It locates other DataNodes that have copies of these blocks and directs I/0 requests to an online DataNode that has a replica of the lost data. It also instructs these DataNodes to copy the "lost" blocks to other nodes whenever possible.

How does adding more data-nodes cause clusters to be ‘unbalanced’?

When new datanodes are added to a cluster, newly created blocks are written to these datanodes from time to time. The existing blocks are not moved to them without using the HDFS Balancer.

What is sparkContext in Apache spark? Can I create RDD without creating sparkContext?

Main entry point for Spark functionality. A SparkContext represents the connection to a Spark cluster, and can be used to create RDDs, accumulators and broadcast variables on that cluster.

Only one SparkContext may be active per JVM. You must stop() the active SparkContext before creating a new one

Yes it can be created without it , options are parallizef collection and from external dataserts

What is the significance of ‘copy-on-write’ in the context of fork() system call?

Copy-on-write finds its main use in sharing the [virtual memory](https://en.wikipedia.org/wiki/Virtual_memory) of [operating system](https://en.wikipedia.org/wiki/Operating_system) [processes](https://en.wikipedia.org/wiki/Computer_process), in the implementation of the [fork system call](https://en.wikipedia.org/wiki/Fork_(system_call)). Typically, the process does not modify any memory and immediately executes a new process, replacing the address space entirely. Thus, it would be wasteful to copy all of the process's memory during a fork, and instead the copy-on-write technique is used. It can be implemented efficiently using the [page table](https://en.wikipedia.org/wiki/Page_table) by marking certain pages of [memory](https://en.wikipedia.org/wiki/Computer_storage) as read-only and keeping a count of the number of references to the page.

What is the main difference between mutex and semaphore

Mutex:

Is a key to a toilet. One person can have the key - occupy the toilet - at the time. When finished, the person gives (frees) the key to the next person in the queue.

Officially: "Mutexes are typically used to serialise access to a section of re-entrant code that cannot be executed concurrently by more than one thread. A mutex object only allows one thread into a controlled section, forcing other threads which attempt to gain access to that section to wait until the first thread has exited from that section." Ref: Symbian Developer Library

(A mutex is really a semaphore with value 1.)

Semaphore:

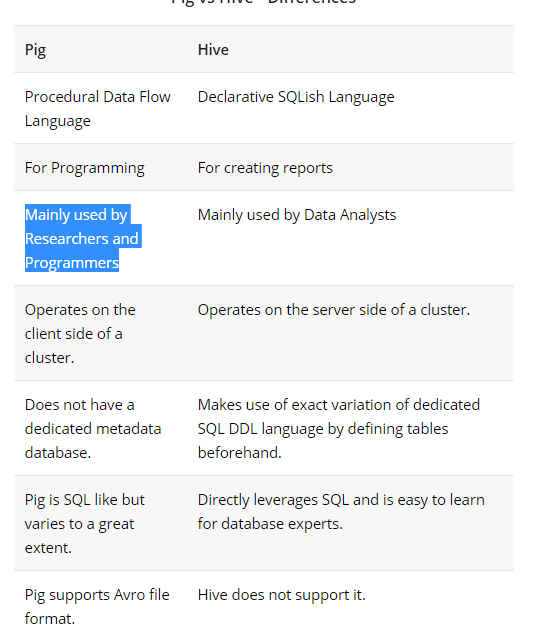
Is the number of free identical toilet keys. Example, say we have four toilets with identical locks and keys. The semaphore count - the count of keys - is set to 4 at beginning (all four toilets are free), then the count value is decremented as people are coming in. If all toilets are full, ie. there are no free keys left, the semaphore count is 0. Now, when eq. one person leaves the toilet, semaphore is increased to 1 (one free key), and given to the next person in the queue.

Officially: "A semaphore restricts the number of simultaneous users of a shared resource up to a maximum number. Threads can request access to the resource (decrementing the semaphore), and can signal that they have finished using the resource (incrementing the semaphore)." Ref: Symbian Developer Library

Mention any 2 forms of IPC in System V

These are message queues, semaphores and shared memory.

Pig vs Hives



DOcker VS VM

Containers are environments that host individual applications using a framework like Docker. Containerized application processes run in isolated environments directly on the host server. There is no hardware emulation required. Nor is there a need to install a complete guest operating system.

In contrast, virtual machines are full operating systems that run on virtual hardware. They are powered by hypervisors like [VMware](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwitsfOj8tHQAhXCi1QKHSg7BO4QFggjMAA&url=http%3A%2F%2Fwww.vmware.com%2F&usg=AFQjCNEvFvErvZHvRukT7j2poj0tjTt3qQ&sig2=AWmmHDhn0l4-muYlEEctPQ&bvm=bv.139782543,d.dGo) and [KVM](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=13&ved=0ahUKEwiGguCx8tHQAhVLhlQKHahXA4MQFghMMAw&url=http%3A%2F%2Fwww.linux-kvm.org%2F&usg=AFQjCNGqy0_3uK5IkZww7hOVPwnXYsw_TA&sig2=NpiJZAisylzUmJ9S5XOiQA&bvm=bv.139782543,d.dGo), which translate computing operations between a guest virtual machine and a host server. The translation process consumes system resources.