**Spark : Cluster Computing with Working Sets**

What does Cluster Computing refer to?

-          Cluster computing refers to the process of sharing computation tasks among multiple machines which works together to act as a single machine

What is a Working Set?

-          A working set is a part of the resources that the current process is using. It is the amount of memory that a process requires in a given time interval

What does RDD stand for?

-          Resilient Distributed dataset

What does it mean when we say an RDD is a collection of objects partitioned across a set of machines?

-          An RDD is distributed, meaning the data is stored across multiple nodes. This helps the data to be resilient. Now, the data itself that is contained in the RDD is not copied across to create resilience, but the lineage of the RDD is, so that the data that is accessed after an action can still be accessed.

Why do we say that MapReduce has an acyclic data flow?

-          MapReduce has a data flow that is not in a cycle or a closed path. The data transformations are similar to that of functional programming, where the process of transforming the data happens programmatically.

Explain the deficiency in using Hive for interactive analysis on datasets. How does Spark alleviate this problem?

-          Hive is more suited for batch processing with large amounts of data. It is not good with streaming data. As such, it runs slower on those interactive analysis processes. Spark, as it runs on RAM, can process streaming data much faster.

 SparkSession vs SparkContext?

* SparkSession
  + Spark SparkContext is an entry point to Spark and defined in org.apache.spark package since 1.x and used to programmatically create Spark RDD, accumulators and broadcast variables on the cluster. Since Spark 2.0 most of the functionalities (methods) available in SparkContext are also available in SparkSession. Its object sc is default available in spark-shell and it can be programmatically created using SparkContext class.
* SparkContext
  + SparkSession introduced in version 2.0 and and is an entry point to underlying Spark functionality in order to programmatically create Spark RDD, DataFrame and DataSet. It’s object ‘spark’ is default available in spark-shell and it can be created programmatically using SparkSession builder pattern.

What is the lineage of an RDD?

-          The list of transformations that the RDD has gone through.

RDDs are lazy and ephemeral. What does this mean?

-          The processing of transformations do not occur until an action is called, and the dataset is discarded from memory after use

What are the 4 ways provided to construct an RDD?

-          Parallelize, from another RDD, from DF/DS, from file

What does it mean to transform an RDD?

-          To transform an RDD is to create a new RDD from an existing one. Transformations are lazy, meaning that the data is not actually transformed until an action is called.

What does it mean to cache an RDD?

-          Caching is an optimization tactic. It saves the data from an RDD action instead of deleting that data right away so that the RDD can be used again on a later transformation or action. This can be helpful if many different transformations and actions are called from a common RDD.

What does it mean to perform a parallel operation on an RDD?

-          Since the RDD is partitioned, operations can be done in parallel, where multiple executors can perform the same action across those partitions

Why does Spark need special tools for shared variables, instead of just declaring, for instance, var counter=0?

-          Because operations happen in parallel across multiple partitions, shared variables need to be able to be accessed in each of those places in memory

What is a broadcast variable?

-          A broadcast variable is a read-only variable that is shared on all nodes. They are immutable and distributed.

What are Persistence Storage Levels in Spark?

- MEMORY\_ONLY, MEMORY\_AND\_DISK, MEMORY\_ONLY\_2, DISK\_ONLY

- Some levels have \_SER, what does this mean?

- Some levels have \_2, what does this mean?

- If the storage level for a persist is MEMORY\_ONLY and there isn't enough memory, what happens?

What is the storage level for .cache()?

* MEMORY\_ONLY

Different ways to repartition?

* .repartition
* .coalesce (to reduce partitions)

What does it mean to run an EMR Step Execution?

* When the job is finished, the cluster terminates. If ran in cluster mode, the job remains idle

What are some benefits to using the cloud?

* Loss prevention, Increased collaboration, Sustainability, Security

What is the Spark History Server?

* Monitoring tool that displays information about completed spark applications

What does it mean to “spill to disk” when executing spark tasks?

* Since spark runs on RAM, if there is not enough memory, the data will spill to disk memory

When during a Job do we need to pay attention to the number of partitions and adjust if necessary?

* If the number of partitions grows to be to high, or if streamed data continuously creates new partitions

What is spark.driver.memory? What about spark.executor.memory?

* The amount of memory that the driver and executor have

What is a Spark Application? Job? Stage? Task?

* Application- main function
* Jobs- Work submitted to Spark. Created when an action occurs
* Stage- Jobs are divided into stages
* Task- Each stage is divided into tasks, the smallest unit of work for Spark

What is the logical plan? The physical plan?

* The logical plan is the tree that represents both the data and schema. They depict the expected output
* The physical plan decides what type of join, sequence of execution of many methods

How many partitions does a single task work on?

- One

What’s the difference between cluster mode and client mode on YARN?

- In cluster mode, the driver runs inside an application master process which is managed by YARN on the cluster and the client can go away after initiation. In client mode, the driver runs in the client process and the application master is only used for requesting resources from YARN

What is an executor? What are executors when we run Spark on YARN?

- Executors are worker nodes’ processes in charge of running individual tasks in a given Spark job.

What is the difference between a DataFrame and a Dataset?

* DF
  + Faster, semantic errors caught at **runtime**, loosely typed (compared to DS), uses Kryo serializer
  + DF = DS[Row]
* DS
  + Slower, semantic errors caught at **compile-time**, strongly typed (case class enforced schema), encoder serializer, cannot be used by Python

What does it mean when DF and DS API’s are unified as of Spark 2.x?

* DF was made to be of type DS[Row]. Many of the same methods can be used for both api’s.