CCA–500 and 505 Exam Sections and Blueprint

Notes: Hadoop ecosystem items are no longer treated separately as their own section and are integrated throughout the exam. Both CCA–500 and CCA–505 share the same proportion of items per section.

1. HDFS (17%)

* Describe the function of HDFS Daemons
* Describe the normal operation of an Apache Hadoop cluster, both in data storage and in data processing.
* Identify current features of computing systems that motivate a system like Apache Hadoop.
* Classify major goals of HDFS Design
* Given a scenario, identify appropriate use case for HDFS Federation
* Identify components and daemon of an HDFS HA-Quorum cluster
* Analyze the role of HDFS security (Kerberos)
* Determine the best data serialization choice for a given scenario
* Describe file read and write paths
* Identify the commands to manipulate files in the Hadoop File System Shell

2. YARN and MapReduce version 2 (MRv2) (17%)

* Understand how upgrading a cluster from Hadoop 1 to Hadoop 2 affects cluster settings
* Understand how to deploy MapReduce v2 (MRv2 / YARN), including all YARN daemons
* Understand basic design strategy for MapReduce v2 (MRv2)
* Determine how YARN handles resource allocations
* Identify the workflow of MapReduce job running on YARN
* Determine which files you must change and how in order to migrate a cluster from MapReduce version 1 (MRv1) to MapReduce version 2 (MRv2) running on YARN.

3. Hadoop Cluster Planning (16%)

* Principal points to consider in choosing the hardware and operating systems to host an Apache Hadoop cluster.
* Analyze the choices in selecting an OS
* Understand kernel tuning and disk swapping
* Given a scenario and workload pattern, identify a hardware configuration appropriate to the scenario
* Given a scenario, determine the ecosystem components your cluster needs to run in order to fulfill the SLA
* Cluster sizing: given a scenario and frequency of execution, identify the specifics for the workload, including CPU, memory, storage, disk I/O
* Disk Sizing and Configuration, including JBOD versus RAID, SANs, virtualization, and disk sizing requirements in a cluster
* Network Topologies: understand network usage in Hadoop (for both HDFS and MapReduce) and propose or identify key network design components for a given scenario

4. Hadoop Cluster Installation and Administration (25%)

* Given a scenario, identify how the cluster will handle disk and machine failures
* Analyze a logging configuration and logging configuration file format
* Understand the basics of Hadoop metrics and cluster health monitoring
* Identify the function and purpose of available tools for cluster monitoring
* Be able to install all the ecoystme components in CDH 5, including (but not limited to): Impala, Flume, Oozie, Hue, Cloudera Manager, Sqoop, Hive, and Pig
* Identify the function and purpose of available tools for managing the Apache Hadoop file system

5. Resource Management (10%)

* Understand the overall design goals of each of Hadoop schedulers
* Given a scenario, determine how the FIFO Scheduler allocates cluster resources
* Given a scenario, determine how the Fair Scheduler allocates cluster resources under YARN
* Given a scenario, determine how the Capacity Scheduler allocates cluster resources

6. Monitoring and Logging (15%)

* Understand the functions and features of Hadoop’s metric collection abilities
* Analyze the NameNode and JobTracker Web UIs
* Understand how to monitor cluster Daemons
* Identify and monitor CPU usage on master nodes
* Describe how to monitor swap and memory allocation on all nodes
* Identify how to view and manage Hadoop’s log files
* Interpret a log file