**HDP Operations: Hadoop Administration I**

**Course Objectives: 5 Days**

* Summarize and enterprise environment including Big Data
* Hadoop and the Hortonworks Data Platform (HDP)
* Manage Ambari Users and Groups
* Manage Hadoop Services
* Use HDFS Storage
* Manage HDFS Storage
* Configure HDFS Storage
* Configure HDFS Transparent Data Encryption
* Configure the YARN Resource Manager
* Submit YARN Jobs
* Configure the YARN Capacity Scheduler
* Add and Remove Cluster Nodes
* Configure HDFS and YARN Rack Awareness
* Configure HDFS and YARN High Availability
* Monitor a Cluster
* Protect a Cluster with Backups
* Hive **(Concept)**
* Pig **(Concept)**
* Hbase **(Concept)**
* Storm **(Concept)**
* Spark Clustering
* kafka
* Data Ingest operations: Flume
  1. Getting Data into HDFS
  2. Ingesting Data from External Sources with Flume
  3. Ingesting Data from Relational Databases with Sqoop
* Data Ingest operations: Kafka **(Concept)**
* Hadoop Security
  1. Hadoop Security System Concepts
  2. Securing a Hadoop Cluster with Kerberos
* Managing and Scheduling Jobs
  1. Managing Running Jobs
  2. Scheduling Hadoop Jobs
  3. Configuring the FairScheduler
  4. Impala Query Scheduling
  5. [Oozie](https://oozie.apache.org/)configuration and job Scheduling
  6. Hadoop schedulers.
  7. Given a scenario, determine how the FIFO Scheduler allocates cluster resources.
  8. Given a scenario, determine how the Fair Scheduler allocates cluster resources.
  9. Given a scenario, determine how the Capacity Scheduler allocates cluster resources
* Cluster Maintenance
  1. Copying Data between Clusters
  2. Adding and Removing Cluster Nodes
  3. Rebalancing the Cluster
  4. Cluster Upgrading
* YARN and MapReduce
  1. What Is MapReduce?
  2. Basic MapReduce Concepts
  3. YARN Cluster Architecture
  4. Resource Allocation
  5. Failure Recovery
  6. Understand basic design strategy for MapReduce v2 (MRv2)
* Planning Your Hadoop Cluster
  1. General Planning Considerations
  2. Choosing the Right Hardware
  3. Network Considerations
  4. Explicitly Including and Excluding Hosts
  5. Principal points to consider in choosing the hardware and operating systems to host an Apache Hadoop cluster.
  6. Given a scenario and workload pattern, identify a hardware configuration appropriate to the scenario
  7. Cluster sizing: given a scenario and frequency of execution, identify the specifics for the workload, including CPU, memory, storage, disk I/O
  8. Disk Sizing and Configuration, including JBOD versus RAID, SANs, virtualization, and disk sizing requirements in a cluster
  9. Network Topologies: understand network usage in Hadoop (for both HDFS and MapReduce) and propose or identify key network design components for a given scenario
  10. Given a scenario, identify how the cluster will handle disk and machine failures.
  11. Understand the basics of Hadoop metrics and cluster health monitoring.
  12. How to increase the external storage in running cluster