Hortonworks Certified Associate (HCA) Exam Objectives

To be fully prepared for the HCA exam, a candidate should be able to perform all of the exam objectives listed below:

Category	Objective	Reference
Data Access	Recognize use cases for Pig, which include: • ETL data pipelines • Researching raw data • Iterative data processing	http://hortonworks.com/apache/pig/
	Understand how Pig executes on a cluster: • Pig runs on YARN as either a MapReduce job or a Tez job. • Pig reads and writes data from the Hadoop Distributed File System (HDFS). • The programming language for the Pig platform is called Pig Latin, which allows you to write a data flow that describes how your data will be transformed and processed (such as filtering, grouping, joining and sorting data).	http://hortonworks.com/apache/pig/#section_2
	Recognize use cases for Hive, which include: • Running SQL queries in Hadoop • Viewing and querying structured and unstructured data by applying a “schema on read” • Allowing data analysts to query data in Hadoop using the familiar SQL syntax	http://hortonworks.com/apache/hive/
	Understand how Hive works on Hadoop: • Hive does not actually store data, but instead allows you to view data in HDFS as SQL-like tables by defining the tables using metadata. • Each Hive table corresponds to a directory in HDFS. • Hive stores the metadata for Hive tables in the Hive metastore, which is a relational database of your choosing. By default, HDP uses MySQL for the Hive metastore. • HiveQL is the query language of Hive and is a subset of the standard SQL • HiveQL queries are converted into Hadoop jobs that run on MapReduce, Tez or Spark	http://hortonworks.com/apache/hive/
	Understand the difference between Hive managed and external tables: • The data for Hive managed tables (also referred to as internal tables) is stored in the Hive warehouse directory in HDFS, which is	https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-ExternalTables

	/apps/hive/warehouse/ by default. • External tables can point to data in any folder in HDFS, which is useful if you already have the data in HDFS that is in a location other than the warehouse directory. • If you DROP a Hive managed table, then its underlying data in HDFS is deleted. • If you DROP an external table, then its underlying data in HDFS is not deleted.	
	Understand use cases for Hive bucketed and partitioned tables: • Bucketed tables allow much more efficient sampling than do non-bucketed tables, and they are also commonly used for performing map-side joins, which are considerably more efficient than reduce-side joins. • Partitioned tables provide a performance benefit by organizing a table's data into subfolders based on a specified column, so that queries with WHERE clauses can save time by only scanning the folders specified by the WHERE condition.	https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL+BucketedTables https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-PartitionedTables
	Understand the purpose of HCatalog: • HCatalog is an extension of Hive that allows other frameworks like Pig and Java MapReduce to access Hive metadata. • HCatalog allows Pig, Hive and Java developers to share a common view of data in HDFS.	http://hortonworks.com/apache/hive/#section_4
	Understand the benefits of Tez: • Tez runs on YARN and executes a DAG (directed acyclic graph) of tasks more efficiently and faster than MapReduce. • Hive and Pig jobs typically run faster on Tez.	http://hortonworks.com/apache/tez/
	Recognize use cases for Storm, which include: • Processing data in real-time by streaming the data through Storm bolts. • Preventing undesirable events from occurring, like a credit card merchant denying a charge in real-time. • Optimizing positive outcomes based on real-time input, like a retail store offering discounts to preferred customers.	http://hortonworks.com/apache/storm/
	Recognize use cases for HBase, which include: • A NoSQL database for Hadoop • Real-time read/write access to large datasets	http://hortonworks.com/apache/hbase/

	• Creating very large tables for storing multi-structured or sparse data. • A financial company that provides stock market ticker data by performing more than thirty thousand reads per second. • A company that provides web security services by maintaining a system that accepts billions of event traces and activity logs from its customers' desktops every day.	
	Explain the purpose of each of the various components of Spark: • Spark SQL: allow Spark applications to execute SQL queries on large datasets written using either a basic SQL syntax or HiveQL. • Spark Streaming: enables scalable, high-throughput, fault-tolerant stream processing of live data streams. • Spark MLlib: Spark's machine learning library, the goal is to provide common algorithms for machine learning that are both scalable and easy to implement. • GraphX: for working with for graphs and performing graph-parallel computation.	http://hortonworks.com/apache/spark/
	Understand how Spark applications execute on YARN: • Cluster mode - the Spark driver runs inside an ApplicationMaster process, which is managed by YARN on the cluster, and the client can go away after initiating the application. • Client mode - the driver runs in the client process. The ApplicationMaster is only used for requesting resources from YARN, and the client application must stay running during the lifetime of the Spark application.	http://spark.apache.org/docs/latest/running-on-yarn.html
	Understand the purpose of Solr: • Solr is a platform for searching data stored in HDFS in Hadoop.	http://hortonworks.com/apache/solr/
Data Management	Understand the Hadoop Distributed File System (HDFS): • HDFS provides the data storage capability of Hadoop. • HDFS is scalable - if you need more storage, you add more nodes to the cluster. • HDFS is fault-tolerant - if a node fails, the data is not lost	http://hortonworks.com/apache/hdfs/

	• The NameNode is the master node that maintains the namespace of the filesystem and sends commands to DataNodes • A Standby NameNode can be configured for NameNode high availability	
	Understand block replication in HDFS: • Large data files are split into blocks that are distributed across the cluster. • The NameNode keeps track of the names of all folders and files, and also the location of the blocks on the DataNodes. • The DataNodes store the blocks of data as instructed by the NameNode.	http://hortonworks.com/apache/hdfs/#section_2
	Understand the purpose of YARN: • Provides the processing component of Hadoop. • The ResourceManager has a scheduler, which is responsible for allocating resources to the various applications running in the cluster, according to constraints such as queue capacities and user limits. • The NodeManagers execute tasks as directed by the ResourceManager. • The ApplicationMaster has responsibility for negotiating appropriate resource containers from the scheduler, tracking their status, and monitoring their progress.	http://hortonworks.com/apache/yarn/
Data Governance and Workflow	Understand the features and capabilities of Falcon: • Falcon simplifies the development and management of data processing pipelines with a higher layer of abstraction, taking the complex coding out of data processing applications by providing out-of-the-box data management services. • Hadoop operators can use the Falcon web UI or the command-line interface (CLI) to create data pipelines, which consist of cluster storage location definitions, dataset feeds, and processing logic.	http://hortonworks.com/apache/falcon/
	Understand the three types of entities in Falcon: • Cluster: Defines where data and processes are stored. • Feed: Defines the datasets to be cleaned and processed. • Process: Consumes feeds, invokes processing logic, and produces further feeds.	http://hortonworks.com/apache/falcon/#section_2

	Understand the purpose of Atlas: • Atlas is designed to exchange metadata with other tools and processes within and outside of the Hadoop stack, thereby enabling platform-agnostic governance controls that effectively address compliance requirements	http://hortonworks.com/apache/atlas/
	Understand the purpose and capabilities of Sqoop: • Sqoop is a tool for transferring data between Hadoop and relational databases. • It works in both directions: data can be offloaded from your EDW into Hadoop for processing, and results can be exported from Hadoop back into your datastore.	http://hortonworks.com/apache/sqoop
	Understand the purpose and capabilities of Flume: • Flume lets Hadoop users ingest high-volume streaming data into HDFS for storage. • Typical sources of these streams are application logs, sensor and machine data, geo-location data and social media.	http://hortonworks.com/apache/flume
	Recognize use cases for Kafka: • Kafka is a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system. • Kafka is often used in place of traditional message brokers like JMS and AMQP because of its higher throughput, reliability and replication.	http://hortonworks.com/apache/kafka/
	Understand the components of Kafka: • Topic: a user-defined category to which messages are published. • Producer: publishes messages to one or more topics. • Consumer: subscribes to topics and processes the published messages. • Broker: manages the persistence and replication of message data.	http://hortonworks.com/apache/kafka/#section_2
	Understand the role of Hortonworks DataFlow (HDF): • HDF is an easy to use, powerful, and reliable system to automate the flow of data between systems. • HDF has a user-friendly drag-and-drop graphical user interface for defining data workflows.	http://hortonworks.com/products/hdf/
Operations	Understand the purpose and capabilities of Cloudbreak:	http://hortonworks.com/apache/cloudbreak

	• Cloudbreak is a tool for provisioning Hadoop clusters on cloud infrastructure such as Amazon Web Services and Microsoft Azure. • Uses Ambari Blueprints to dynamically configure and provision HDP clusters in the cloud. • You pick an Ambari Blueprint, then pick a cloud provider, and Cloudbreak does the rest.	
	Understand the purpose of Ambari: • Ambari is a web application for provisioning, managing, monitoring and securing Hadoop clusters. • System Administrators use Ambari to install clusters, install and remove HDP services, ensure the various components of HDP are up and running, and collect metrics to analyze the performance and efficiency of the cluster.	http://hortonworks.com/apache/ambari
	Understand the purpose of Ambari Views: • A view is an application that is deployed into the Ambari container. • A view is a way of extending Ambari that allows 3rd parties to plug in new resource types along with the APIs, providers and UI to support them. • For example, the Hive View allows users to write & execute SQL queries on the cluster. • The Pig View allows the writing and running of Pig scripts. • The Capacity Scheduler View allows users to define and configure YARN queues.	http://hortonworks.com/apache/ambari/#section_4
	Understand the role of ZooKeeper in an HDP cluster, which includes: • Providing a distributed configuration service. • Providing a synchronization service. • Providing a naming registry for distributed systems. • For example, ZooKeeper plays a key role in NameNode and YARN High Availability.	http://hortonworks.com/apache/zookeeper
	Understand the purpose and capabilities of Oozie: • Oozie is a Web application used to schedule Hadoop jobs. • An Oozie Workflow is a sequence of actions to be performed, where the actions could be a Pig job, Hive query, MapReduce job and so on.	http://hortonworks.com/apache/oozie

	• An Oozie Coordinator job is used to trigger when a Workflow executes.	
Security	Understand the purpose and capabilities of Ranger: • Ranger offers a centralized security framework to manage fine-grained access control over Hadoop data access components like Hive and HBase. • Using the Ranger console, security administrators can easily manage policies for access to files, folders, databases, tables, or columns.	http://hortonworks.com/apache/ranger
	Understand the purpose and capabilities of Knox: • Knox provides perimeter security for Hadoop clusters. • The Knox Gateway provides a single access point for all REST interactions with Apache Hadoop clusters. Knox can work directly with Kerberos for authentication and authorization of users.	http://hortonworks.com/apache/knox-gateway/