Required Skills

Data Ingest

The skills to transfer data between external systems and your cluster. This includes the following:

* Import data from a MySQL database into HDFS using Sqoop
* Export data to a MySQL database from HDFS using Sqoop
* Change the delimiter and file format of data during import using Sqoop
* Ingest real-time and near-real-time streaming data into HDFS
* Process streaming data as it is loaded onto the cluster
* Load data into and out of HDFS using the Hadoop File System commands

Transform, Stage, and Store

Convert a set of data values in a given format stored in HDFS into new data values or a new data format and write them into HDFS.

* Load RDD data from HDFS for use in Spark applications
* Write the results from an RDD back into HDFS using Spark
* Read and write files in a variety of file formats
* Perform standard extract, transform, load (ETL) processes on data

Data Analysis

Use Spark SQL to interact with the metastore programmatically in your applications. Generate reports by using queries against loaded data.

* Use metastore tables as an input source or an output sink for Spark applications
* Understand the fundamentals of querying datasets in Spark
* Filter data using Spark
* Write queries that calculate aggregate statistics
* Join disparate datasets using Spark
* Produce ranked or sorted data