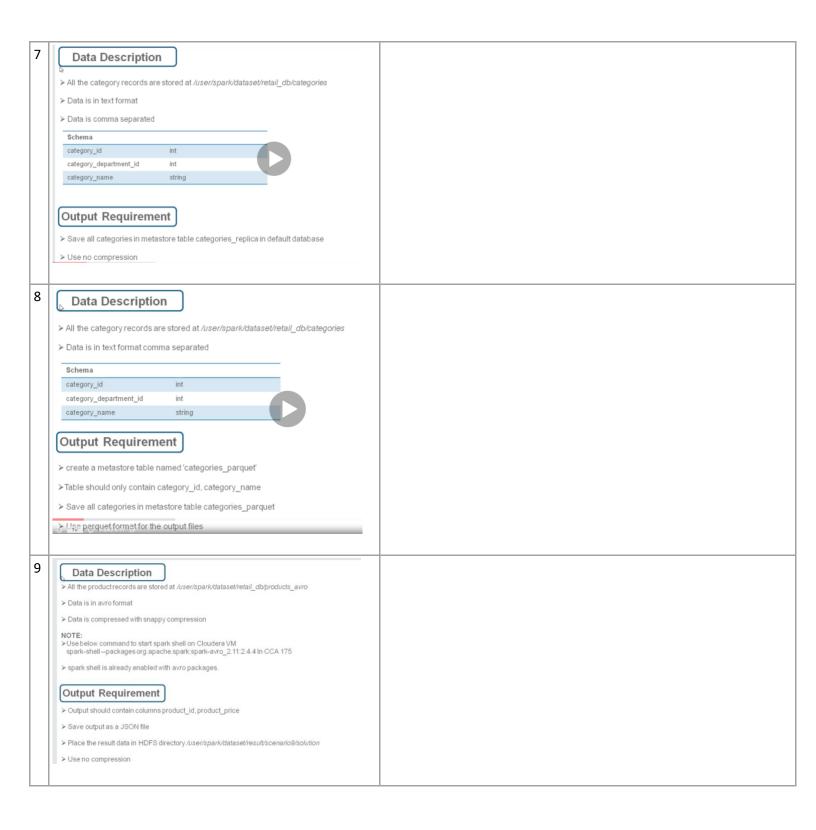


## **Data Description** Output Requirement > Convert data into tab delimited file > All the categories records are stored at /user/spark/dataset/retail\_db/categories > Use text format for the output files > Place the result data in HDFS directory /user/spark/dataset/result/scenario4/solution > Data is in text format > Compress the output using Iz4 compression > Data is comma separated Schema category\_id category\_department\_id int category\_name string 5 **Data Description** > All the product records are stored at /user/spark/dataset/retail\_db/products\_avro > Data is in avro format > Data is compressed with snappy compression Output Requirement > Output should only contain the products with price greator than 1000.0 $\succ$ Use parquet format for the output files > Place the result data in HDFS directory /user/spark/dataset/result/scenario5/solution > Compress the output using snappy compression 6 **Data Description** > Get data from metastore table named "orders" > Table is present in the database "default" Output Requirement > Fetch orders from Jan-2013 to Dec-2013 > Use parquet format for the output files > Place the result data in HDFS directory /user/spark/dataset/result/scenario6/solution > Compress the output using Gzip compression



1 0	Data Description  > All the category records at  > Data is in text format com-	are stored at /user/spark/dataset/re	etail_db/categories								
	Schema	200 20 (00 20 20 20 20 20 20 20 20 20 20 20 20 2									
	category_id	int									
	category_department_id	int									
	category_name	string									
		named 'categories_partitioned' tastore table categories_partitione	d								
				T							