Najpierw przekonfiguruj yarna:

1. Na cloudera managerze wejdź w konfigurację yarna.
2. Wpisz w wyszukiwarkę: yarn.nodemanager.resource.cpu-vcores i ustaw ten parametr na 2.
3. yarn.app.mapreduce.am.resource.mb: 512MiB
4. mapreduce.map.memory.mb: 512MiB
5. mapreduce.reduce.memory.mb: 512MiB
6. yarn.scheduler.minimum-allocation-mb: 512MiB
7. yarn.scheduler.increment-allocation-mb: 512MiB
8. yarn.scheduler.maximum-allocation-mb: 512MiB
9. mapreduce.map.java.opts.max.heap: 400 MiB
10. mapreduce.reduce.java.opts.max.heap: 400 MiB
11. 'ApplicationMaster Java Maximum Heap Size': 400 MiB
12. Kliknij save changes
13. Kliknij w logo cloudera managera w lewym górnym rogu. Przy yarnie pojawiła się ikonka z tooltipem: 'Stale configuration ...'. kliknij ją
14. Kliknij przycisk: 'Restart stale services'

Job hadoop streaming na oozie:

1. wrzuć mapper.py i reducer.py na hdfsa
2. rozwiń na hue zakładkę 'Workflows' i następnie kliknij w 'Editors'
3. kliknij po prawej stronie przycisk 'create'
4. Przeciągnij akcję streaming do workflowu
5. W pole Mapper wpisz 'mapper.py', a w pole Reducer wpisz 'reducer.py'
6. Kliknij 'Add'
7. Kliknij w przycisk 'FILES+' i znajdź plik 'mapper.py', następnie kliknij jeszcze raz i znajdź plik 'reducer.py'
8. Kliknij w prawym górnym rogu akcji w przycisk ustawień
9. Kliknij w przycisk 'PROPERTIES+' i wpisz w pierwsze pole 'mapred.input.dir', a w drugim podaj ścieżkę do katalogu / pliku wejściowego
10. Kliknij jeszcze raz i dodaj property 'mapred.output.dir' z namiarami na katalog wyjściowy
11. Zapisz workflow klikając save w prawym górnym rogu

Job mapreduce:

1. wrzuć jara z jobem na hdfsa
2. dodaj akcję 'Java program' do workflowu
3. w polu 'Jar name' znajdź jara z jobem na hdfsie
4. w polu 'Main class' wpisz nazwę klasy razem z pakietem
5. Kliknij 2 razy w 'ARGUMENTS+'. W pierwszym polu wpisz plik/katalog wejściowy, a w drugim wyjściowy
6. Zapisz workflow klikając save w prawym górnym rogu

Job sqoop

1. wrzuć 'mysql-connector-java-5.1.39-bin.jar' na hdfsa
2. dodaj akcję 'sqoop1' do workflowu
3. w polu Śqoop command"wpisz komendę do importu sqoop (bez polecenia sqoop) i kliknij add
4. Kliknij w prawym górnym rogu akcji w przycisk ustawień
5. kliknij na 'ARCHIVES+'
6. znajdź na 'hdfsie mysql-connector-java-5.1.39-bin.jar'
7. Zapisz workflow klikając save w prawym górnym rogu

Job hive

1. stwórz skrypt 'nazwa.sql' zawierający komenty sqlowe do wykonania
2. umieść skrypt 'nazwa.sql' na hdfsie
3. dodaj akcję 'HiveServer2' do workflowu
4. podaj namiary na skrypt 'nazwa.sql' i kliknij 'add'
5. Zapisz workflow klikając save w prawym górnym rogu

Koordynator

1. Kliknij na dropdown 'Workflows'
2. Najedź na 'Editors'
3. Kliknij na 'Coordinators'
4. Kliknij przycisk 'Create' po prawej
5. Kliknij 'Choose a workflow' i wybierz workflow
6. Wybierz jak często job ma się wykonywać
7. kliknij save, żeby zapisać koordynatora

Zadania:

1. Utwórz workflow z joba streaming sliczającego słowa i joba mapreduce sortującego po liczbie wystąpień
2. Utwórz workflow ściągający dane ze sqoopa, ładujący je do tabeli na hivie i wyliczający tabelę wynikową będącą połączeniem apache_logs z tabelą ip_name