Oozie

Hadoop Streaming

1. wrzuć mapper.py i reducer.py na hdfsa
2. rozwiń na hue dropdown menu przy przycisku 'Query', wejdź w menu 'Scheduler' i kliknij 'Workflow'
3. kliknij w dropdown 'DOCUMENTS' i wybierz 'Actions'
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

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

Sqoop

1. wrzuć 'mysql-connector-java-5.1.46-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-5.1.46-bin.jar'
7. Zapisz workflow klikając save w prawym górnym rogu

Hive

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

Koordynator

1. rozwiń dropdown menu przy przycisku 'Query', wejdź w menu 'Scheduler' i kliknij 'Schedule'
2. Kliknij na dropdown 'Choose a workflow...' i wybierz workflow, który chciałbyś zaschedulować.
3. Wybierz jak często job ma się wykonywać. Możesz kliknąć w Options i zaznaczyć Ądvanced syntax", żeby wpisać wyrażenie crona.
4. Wybierz zakres dat w jakim scheduler ma się wykonywać
5. kliknij save, żeby zapisać koordynatora

Dodanie parametru do joba

```
${nazwa_parametru}
```

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 'owners' na hivie i wyliczający tabelę wynikową będącą połączeniem tabeli 'tranfers' z tabelą 'owners'