Anhang A – User Guide

MapReduce-Job

Ziel

PDF-Dateien aus dem HDFS in Elasticsearch indexieren.

Vorrausetzung

Schreibberechtigung für Elasticsearch und Leseberechtigung für das HDFS.

Ausführung

Um den MapReduce-Job auszuführen, muss der Inhalt des Shell-Skripts wie folgt aussehen:

- #! /bin/sh
- hadoop jar <path to jar> <elastic username> <elastic password> <host:port> <index/type> <file location in HDFS>

Nachdem das Skript gespeichert wurde, kann das Skript ausgeführt werden, indem der Skript-Name wie folgt angegeben wird:

- bash <Skriptbezeichnung>

Ausführungsbeispiel:

Angenommen die oben beschriebenen Variablen sind folgende:

- path to jar: jamshedi/ElasticIndexing.jar,
- package.mainclass: com.tiggs.core.PDFDriver,
- elastic username: jamshedi,
- elastic password: 123456,
- host:port: 12.3.4.567:9200,
- index/type: hadoop/pdf/,
- file location in HDFS: hdfs://12.3.4.567:9000/user/jamshedi/customer/Example.pdf.

Dann sieht der Inhalt des Skripts wie folgt aus:

- #! /bin/sh
- hadoop jar jamshedi/ElasticIndexing.jar com.tiggs.core.PDFDriver 12.3.4.567:9200 hadoop/pdf/ hdfs://12.3.4.567:9000/user/jamshedi/Example.pdf

Hinweis

Wenn mehre Pdfs in einem Verzechnis ausgewählt werden sollen, dann kann dies durch Einsatz von Wildcards ermöglicht werden.

- hdfs://12.3.4.567:9000/user/jamshedi/customer/.*pdf

Softwarelösung

Vorrausetzung

Schreibberechtigung für MongoDB und Leseberechtigung für Elasticsearch.

Ausführung

- Java -cp <path to jar> <package.mainclass> <path to property file>

Ausfürhungsbeispiel:

Angenommen die oben beschriebenen Variablen sind folgende:

- path to jar: jamshedi/ElasticIndexing.jar,
- package.mainclass: com.tiggs.core.Main,
- path to property file: C:\Users\Siyar\Desktop\ELCOMP.properties.

Dann sieht der Inhalt der Batch-Datei wie folgt aus:

- java -cp C:\Users\Siyar\Desktop\ELComp-1.0-SNAPSHOT.jar com.tiggs.core.Main C:\Users\Siyar\Desktop\ELCOMP.properties
- pause

Beschreibung der Key/Value-Paare innerhalb der Property-Datei

Property	Beschreibung
ElsUser	Username des Elasticsearch-Benutzers.
ElsPassword	Passwort des Elasticsearch-Benutzers.
ElsHost	Host der Elasticsearch-Instanz.
ElsHostPort	Port zum zugehörigen Host der Elasticsearch-Instanz.
MongoURI	Verbindungs-String zur MongoDB-Instanz.
ElsIndex	Der in Elasticsearch durchzusuchende Index.
ElsFiledname	Das Feld zum zugehörigen Index das durchsucht werden soll.
ElsFieldvalue	Der Wert des durchzusuchenden Feldes.

ElsSearchStartIndex	Ab welchem Index das Ergebnis ausgegeben werden soll. Bei Eingabe von 10 werden die ersten 10 Ergebnisse nicht angezeigt.
ElsSearchMaxResponseSize	Bis zu wie viele Treffer ein Ergebnis maximal haben darf.
MongoDataBaseName	Bezeichnung der MongoDB-Datenbank innerhalb der MongoDB-Instanz.
MongoCollectionName	Bezeichnung der MongoDB-Collection innerhalb der MongoDB-Datenbank.
ElsDefaultResponseStructure	Wenn dieses Feld als *true* gesetzt ist, brauchen die nächsten 4 Felder nicht ausgefüllt zu werden. Die Such-API von Elasticsearch hat zum jetzigen Zeitpunkt eine bestimmte Antwortstruktur, die sich jedoch durch Aktualisierungen ändern kann. Um den Code in diesem Fall nicht ändern zu müssen, können mit den nächsten 4 Feldern neue *keys* angegeben werden.
ElsRootObjectKey	Das erste JSON-Objekt der Search-API Response.
ElsFirstArrayKey	Das erste Array innerhalb des JSONObjekts.
ElsObjectInArrayKey	Das erste Objekt innerhalb des zuvor beschriebenen Arrays.
ElsDesiredValueKey	Der gewünschte Key, der den Dateiinhalt der PDF-Dateien innerhalb des Objekts enthält.

Tabelle 1: Beschreibung Property Key/Value

Property-Datei Beispiel

```
ElsUser=jamshedi
ElsPassword=DywG7Z9Xfp8jcT0HtvQV
ElsHost=localhost
ElsHostPort=9200
MongoURI=mongodb+srv://Admin:Admin@testcluster-xys8u
ElsIndex=hadoop
ElsFiledname=content
ElsFieldvalue=Jamshedi
ElsSearchStartIndex=0
ElsSearchMaxResponseSize=10000
MongoDataBaseName=ElasticDB
```

```
MongoCollectionName=ElasticHadoop
# when set to false Els key properties have to be set
ElsDefaultDocumentStructure=true
ElsRootObjectKey=null
ElsFirstArrayKey=null
ElsObjectInArrayKey=null
ElsDesiredValueKey=null
```