И.А. Еремин
Муромский институт (филиал) Владимирского государственного университета
Россия, 602264, Владимирская область, г. Муром, ул. Орловская, д.23
E-mail: ilya071294@gmail.com

Разработка утилиты миграции БД с Oracle на Ред База Данных

В настоящее время на государственном уровне отдается приоритет отечественному программному обеспечению, поэтому часто возникают задачи по переносу решений, использующих иностранные разработки. В связи с необходимостью миграции БД с Oracle на Ред База Данных, появилась идея создать утилиту, которая позволит автоматизировать большинство операций преобразования.

Общий принцип работы утилиты предполагался следующий:

1) выполнение синтаксического анализа скрипта создания базы данных Oracle;

2) построение синтаксического дерева по полученным данным;

3) обход дерева с преобразованием отдельных узлов, в результате чего будет сформирован скрипт создания базы данных на РБД.

Процесс группировки символов исходного текста в слова или лексемы называется лексическим анализом, а программа, выполняющая его – лексером. Лексер группирует лексемы по типам (например, целые числа, идентификаторы, вещественные числа и т.д.). Каждая лексема описывается по крайней мере двумя свойствами: типом лексемы и соответствующим этому типу текстом из исходного файла [1].

Синтаксический анализатор (парсер) принимает на вход поток лексем и выполняет распознавание структуры выражений. В результате строится синтаксическое дерево (дерево разбора), которое наглядно показывает, каким образом выполнен разбор исходного текста.

Написание модулей, разбирающих эти данные, вручную является очень трудоемкой задачей и не всегда гарантирует хороший результат. Существуют специальные средства, позволяющие облегчить создание парсеров – генераторы синтаксических анализаторов. Подобные средства оказываются незаменимыми, если необходимо использовать в работе программы данные какого-нибудь сложного формата [2].

Для создания синтаксического анализатора грамматики Oracle был выбран генератор анализаторов для формальных языков ANTLR4 [3]. ANTLR генерирует код лексера и парсера, используя файл с описанием грамматики разбираемого языка, который состоит из лексических и синтаксических правил, описанных в расширенной форме Бэкуса – Наура [4].

Выбор ANTLR4 обоснован тем, что в нем есть возможность по описанной грамматике сгенерировать набор триггеров, каждый из которых соответствует какому-либо синтаксическому правилу. При обходе дерева разбора производится вызов того триггера, который соответствует текущему типу узла. Внутри триггера можно выделять из общего потока лексем только те, которые принадлежат текущему синтаксическому правилу. Над полученными лексемами можно выполнять операции вставки, замены и удаления.

В процессе написания программы с использованием подхода с триггерами возникли проблемы, требующие решения:

1) нет контроля над форматированием выходного скрипта;

2) невозможно контролировать порядок вставки новых лексем;

3) отсутствует возможность переупорядочивания SQL-операторов.

Литература

1. Terence Parr. The Definitive ANTLR Reference, The Pragmatic Bookshelf, 2012. –322 с.
2. Написание парсеров с помощью ANTLR // URL: http://club.shelek.ru/viewart.php?id=39
3. ANTLR // URL: http://www.antlr.org/
4. Форма Бэкуса – Наура // URL: https://ru.wikipedia.org/wiki/Форма_Бэкуса_—_Наура.