Using R Efficiently with Large Databases

Dr. Michael Wurst, IBM Corporation
Architect – R/Python Database Integration, In-Database Analytics
Patterns of Database Integration

Pulling Data into R

Pros: Only limited by data size, work with R “as usual”
Cons: Data size is limited, often mix of R and SQL code that is hard to read, if not using a specialized driver, loading data from R can be very slow, non-parallel data access

RODBC, RJDBC, custom packages (based on DBI)
Translate R code into SQL (using proxy objects), either imitating the behavior of R methods and functions or creating a set of explicit functions for transforming data (e.g. dplyr)
Patterns of Database Integration

SQL Push-Down

```r
idf <- ida.data.frame('HUGETABLE')
head(idf[,c('V1','V2')],3)

SELECT V1,V2 FROM HUGETABLE
FETCH FIRST 3 ROWS ONLY
```

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.3</td>
</tr>
<tr>
<td>2</td>
<td>10.9</td>
</tr>
<tr>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>4.7</td>
</tr>
</tbody>
</table>
Patterns of Database Integration

SQL Push-Down

\[ \text{idf} \leftarrow \text{ida.data.frame('HUGETABLE')} \]
\[ \text{idaLm(AGE~INCOME, idf)} \]
\[ \text{SELECT SUM(X1*X2), [..]} \]
\[ \text{[..]} \]
Patterns of Database Integration

SQL Push-Down

**Pros:** No need to know/write SQL, profits from scalability/indexing of the data warehouse (e.g. columnar storage)

**Cons:** Usually only a subset of R functionality can be pushed down in this way
Patterns of Database Integration

Running R code In-Database

Pros: Can execute almost any R code, call R code from SQL

Cons: Debugging, workload management, security are more complex, most R packages do not scale out-of-the-box

\[
f <- \text{function}(x) \{ x[[2]] + 6 \}
\]

\[
nzdf <- \text{nz.data.frame('HUGE TABLE')}
\]

\[
r <- \text{nzApply}(nzdf,f,'outtab')
\]

\[
<\text{serialized R code}>
\]

\[
nz.data.frame('outtab')
\]
Summary

• Each pattern has some particular benefits and drawbacks, you might need all of them at some point.

• There is still no actual standard, especially when exploiting features specific to some database management systems.

• Technologies like Apache SparkR will also influence how we work with Databases.