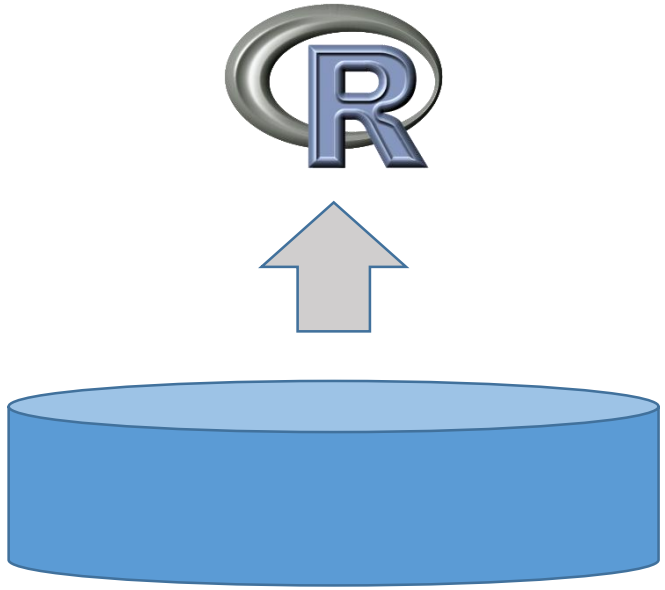
Using R Efficiently with Large Databases

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

Patterns of Database Integration

Pulling Data into R

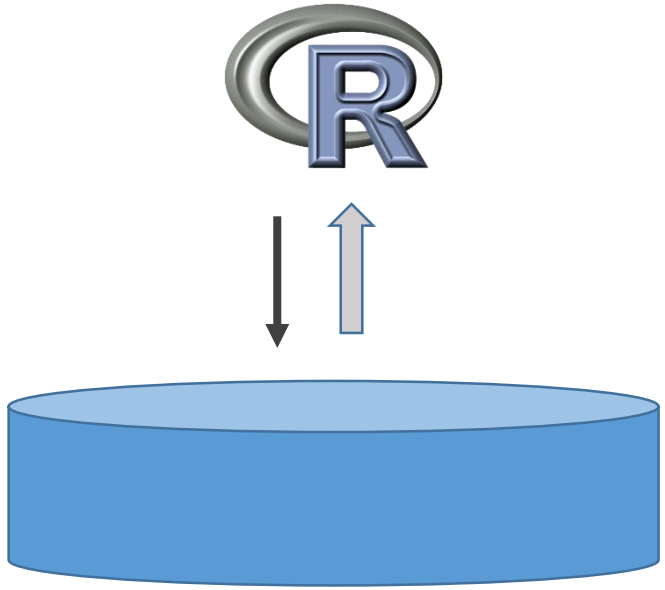


RODBC, RJDBC,
custom packages (based on DBI)

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

Patterns of Database Integration

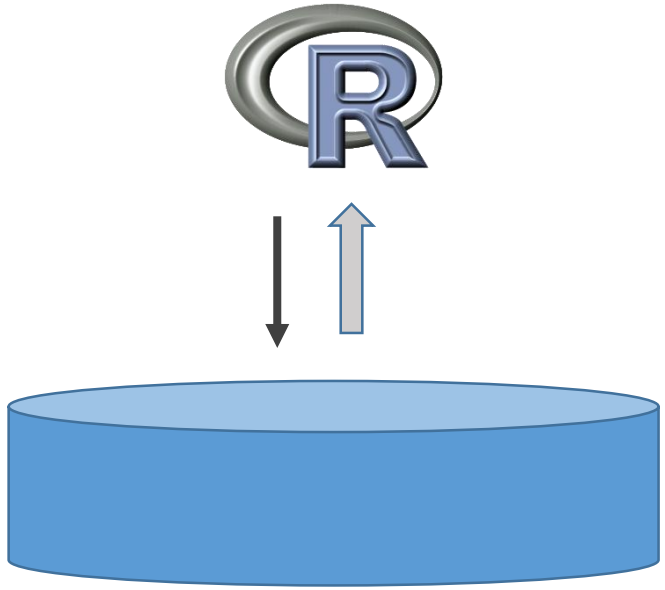
SQL Push-Down



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



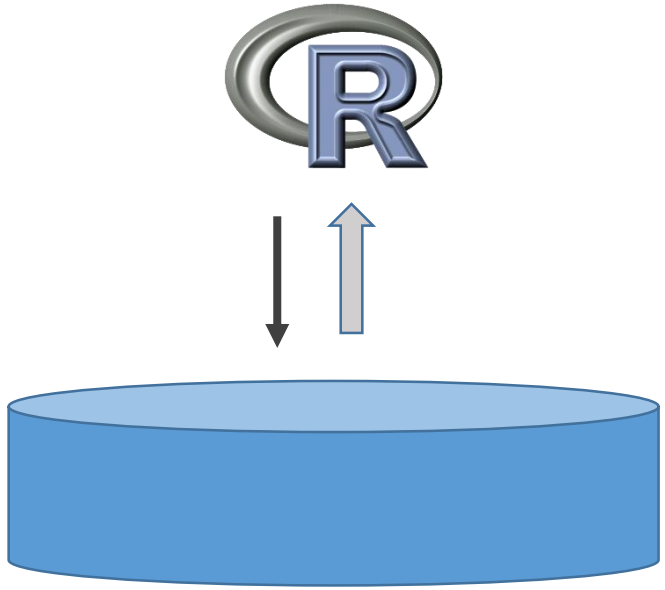
```
idf <- ida.data.frame('HUGETABLE')
head(idf[,c('V1','V2')],3)
```

SELECT V1,V2 FROM HUGETABLE
FETCH FIRST 3 ROWS ONLY

	V1	V2
1	5.3	3.5
2	10.9	3.0
3	1.2	4.7

Patterns of Database Integration

SQL Push-Down



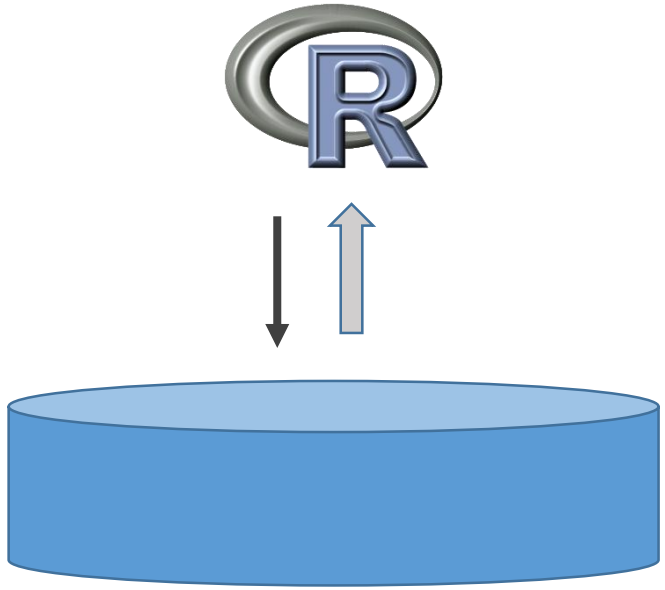
```
idf <- ida.data.frame('HUGETABLE')
idaLm(AGE~INCOME,idf)
```

[..]
SELECT SUM(X1*X2), [..]
[..]

lm-like object

Patterns of Database Integration

SQL Push-Down

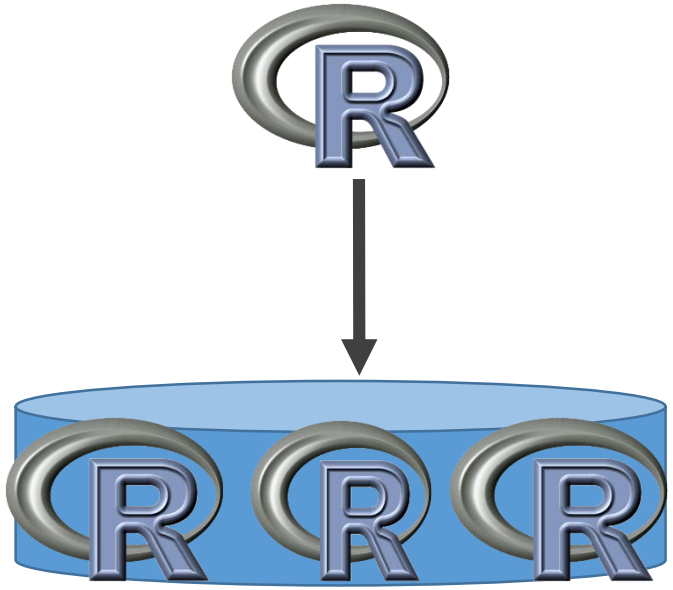


Pros: No need to know/write SQL, profits from scalability/indexing of the data warehouse (e.g. columnar storage)

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

Patterns of Database Integration

Running R code In-Database



```
f <- function(x) {x[[2]]+6}
nzdf <- nz.data.frame('HUGETABLE')
r <- nzApply(nzdf,f,'outtab')
```

<serialized R code>

nz.data.frame('outtab')

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

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.