Better Living Via Automation

Introduction to RODBC & SQL

ATL R Users Group

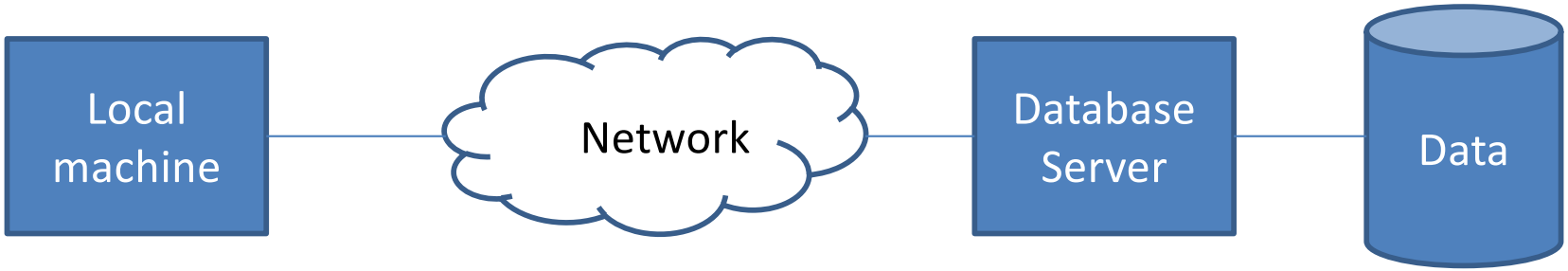
Realities of Working In Industry

- Data Collection often organized around Processes
  - Order-to-cash, Procure-to-Pay, Inventory, BOM, etc..
- Roles often focused on deep analysis of specific domain
  - Exploratory analysis
  - Modeling Building / Validation
  - Tracking / Change Management
- **The Good:** data often aggregated, can do CSV download
  - Corporate data warehouse (ERP offshoot => dedicated box)
  - Departmental Systems (MS Access) / Web Platforms
- **The Bad:** CSV generation often manual…
- **The Ugly:** Weaving together a dozen views for a weekly update…
  - "Manager of Spinning Blue Circles"… grrr…
- Is there a faster way?

RODBC

- ODBC - Open Database Connectivity
- Industry standard API for talking to databases
- Supported by most major databases since ‘90’s
  - Database agnostic
  - Use SQL (also near standard) to make requests
- Platform specific drivers handle implementation details
- RODBC – library for tapping into this API

Why Use ODBC?



- All R operations take place in local machine memory
- Corporate data stores often contain hundreds of gigabytes to terabytes of information, exceeding capacity of desktops / laptops
- ODBC enables one to pre-process / aggregate data on a remote server, sending the query results to the local machine (versus moving all the data first and then processing it locally, ***as long as the remote database supports the operations defined in an SQL statement passed to it***

Scripting Database Access

- install.packages("RODBC")
- library("RODBC")

connect to database. (Access specific issue: 32 bit vs. 64 bit drivers)

- conn <- odbcConnectAccess2007("C:/Users/aioffers/Documents/R Meetup/rodbc/test.accdb")

create SQL string, use to query database

- sqlresult <- SQLQuery("Select * from parts")

	ID	PartNumber	Manufacturer	Distributor	CaseWeight	CaseQuantity	Material	Demand
1	8	1	BottleWerkz	BottleShop	25	100	HDPE	500
2	9	2	BottleWerkz	BottleShop	35	150	HDPE	5000
3	10	3	BottleMaker	BottleShop	533	2500	PET	2500
4	11	4	BottleForge	BottleBuds	200	2000	PET	10000
5	12	5	LastGlass	BottleBuds	400	1000	Glass	7500
6	13	6	KanDo	BottleShop	500	1250	Steel	4000

Some Basic SQL – Single Table

```
sqlresult <- sqlQuery(conn,
    "Select material,                                          Specific field
    sum(demand*caseweight) as weight                           Calculated Field
    from parts                                                 Table name
    group by material                                          Aggregation
    order by sum(demand*caseweight) desc")                     Sort order (Reverse)
```

	material	weight
1	Glass	3000000
2	HDPE	187500
3	PET	3332500
4	Steel	2000000

Some Basic SQL – Basic Join

```
sqlresult <- sqlQuery(conn,
" SELECT parts.Material,
    Sum([parts]![Demand]*[parts]![CaseWeight]) AS Weight,
    Sum(([parts]![Demand]*[parts]![CaseWeight])*[Freight]![F
    reightRate]) AS Freight
FROM Freight INNER JOIN parts ON Freight.Vendor =
    parts.Distributor
GROUP BY parts.Material;")
```

Join Tables Using Like Fields

	Material	Weight	Freight
1	Glass	3000000	270000
2	HDPE	187500	26250
3	PET	3332500	366550
4	Steel	2000000	280000

SQL – More Explorations…

- Core statements:
  - Create (“Insert”)
  - Read (“Select”)
  - Update
  - Delete
- Access Rights / Roles – Managed At Table Level
  - Analysts rarely have write access to production…. (and no, you don’t want it…)
- If the query runs slow, look at the complexity of joins and # records
- Window Functions – Oracle, SQL Server
  - Ranks, Ranges, Running Totals…
- The nicer (Paid) databases have embedded statistical libraries
  - Push heavy calculations to the database server

Alternative Packages

You might want to mention something about security / credentials here

- DBI
- Odbc
- RMySQL / RMariaDB
- Roracle
- RSQLite
- (and many more…)

- Generally takes an afternoon to solve for the connection details at a specific organization

- Other Tips
  - Automate “data conversion”: use meta-data about query results to trigger conversions to appropriate R data types
  - Cron (and other schedulers)… can easily run at night…
  - Database writes - remember to commit..
  - Prototype / Explore in R, push SQL views to web-based reports or Shiny Apps…