R – A Data Analyst Toolbox

Tri Nguyen

Feb 2015
Toronto Hadoop User Group

Target Audience

- Data Analysts
- DB Developers
- Maths
- Data Sciences

Typical Data Challenge

You Have

- New Dataset to investigate
- Messy, Missing Data

You Want

- Answers to concret business questions
- Extract Features
- Detect Trends, Relation, Outliers
- Predictions

Traditional Tools

- ETL
- SQL, Tableau, Excel
- Datawarehouse, OLAP

Possible Inconveniences

- Long design time
- Could miss some insights:
  relation, quantify noise or errors

Design Time

Most Polluted Cities in Canada

- Parse 10MB Fixed Length File
- Calculate Average Ozone per City
- Sort descending

	Exec Duration	Slow Factor	Design Time
SQL 2012 R2 (*)	**6 secs**		4 hours
Pig (**)	154 secs	x25	2 hours
Hive (**)	100 secs	x16	4 hours
Java MapReduce (**)	49 secs	x8	30 hours

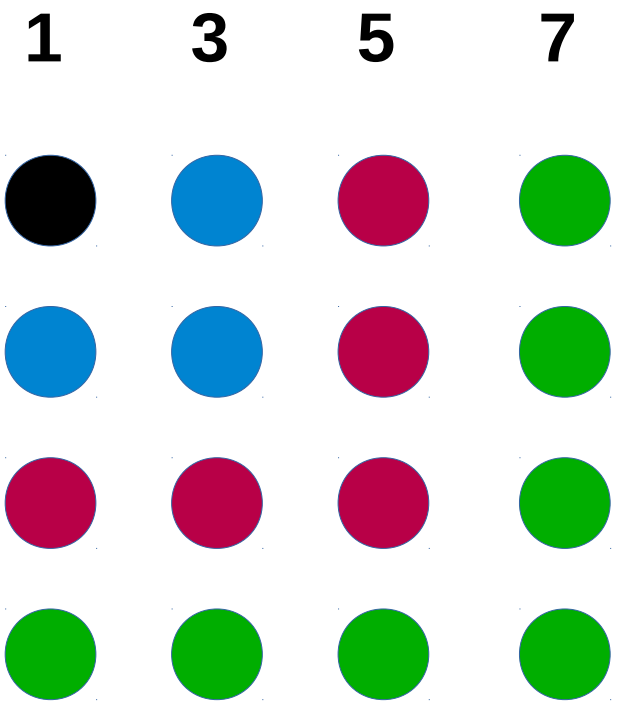
Using R

- 1 hour max in design time
- Graphics & More Calculations

Demo #1 (Maths)

- Starting from 1
- The sum of all consecutive odd numbers = perfect square
- $1 + 3 + 5 = 9 = 3^2$

Proof #1



Demo #2 (Engineering)

Where is the Plane?

R - Features

- ETL
  - Read/Write data stream, various sources & format
  - Cleanup & Transform
- Exploratory Data Analysis
- Statistics
- Machine Learning
- Presentation
  - Graphics
  - Publishing: HTML, PDF, Slides
  - interactive query like Tableau

Query #1 (select)

```
SELECT Continent, Country, CurrencyCode, Population, GDP2013,
   GDPperCapita AS (1000000 * GDP2013/Population)
FROM dtGDP
WHERE Continent="Europe" AND GDP2013 > 1000000
ORDER BY GDPperCapita DESC
```

R (using dplyr)

```
dtGDP %>%
  filter(Continent=="Europe" & GDP2013 > 1E6) %>%
  transmute(Continent, Country, CurrencyCode, Population, GDP2013,
     GDPperCapita = as.integer(1E6 * GDP2013/Population)) %>%
  arrange(desc(GDPperCapita))
```

R (data.table)

```
dtGDP[ Continent=="Europe" & GDP2013 > 1E6,
  .(Continent, Country, CurrencyCode, Population, GDP2013)
  ] [ , GDPperCapita := as.integer(1E6 * GDP2013/Population)
  ] [order(-GDPperCapita)]
```

Query #2 (Aggregate)

```
SELECT Continent, AVG(GPD) AS AvgGDP, COUNT(*) AS CountryCount
FROM dtGDP
GROUP BY Continent
ORDER BY AvgGDP DESC
```

R (using dplyr)

```
dtGDP %>%
   group_by(Continent) %>%
   summarize(CountryCount=n(), AvgGDP=mean(GDP2013)) %>%
   arrange(desc(AvgGDP))
```

R (data.table)

```
dtGDP[, .(AvgGDP = mean(GDP2013), CountryCount=.N), by=Continent
   ] [, .(Continent, CountryCount, AvgGDP)] [order(-AvgGDP)]
```

Query #3 (Join)

```
SELECT TOP 10 C.Country, C.Continent, ECO.GDPperCapita, ECO.PopuDensity
    FROM dtCountry C
INNER JOIN dtEconomy ECO ON ECO.CountryCode = C.CountryCode
ORDER BY Country
```

R (using dplyr)

```
dtCountry %>%
  inner_join(dtEconomy, by="CountryCode") %>%
  select(Country, Continent, GDPperCapita, PopuDensity) %>%
  arrange(Country) %>% filter(row_number() <= 10)
```

R (data.table)

```
dtCountry[ dtEconomy,
  .(Country, Continent, GDPperCapita, PopuDensity)
  ] [order(Country)] [1:10]
```

Demo #3 (Datathon)

http://rpubs.com/TriNguyen/DatathonDec2014
https://github.com/NeuroNex/Samara/tree/master/ElectionCanada_AnalysisTeam1

- Election Canada donations 2004-2013
- Web Extracts -> CSV, 1.6 million rows
- Serious clean up

Questions

- Tax Credits diverted into political parties?
- Contribution $ per Capita, per Party?
- Donation to 1 or N party?
- Lump sum or periodic donation?

R as a Language?

- Oriented “single mission” task
  (not suitable to build application)
- Mix of Procedural, Object, Functional
- Constantly evolving
  Experts sometimes stuck with legacy code

Best Learning Approaches:

- Use R to solve a concrete business problem
- Think in R