
O'REILLY®

Up Your R Game

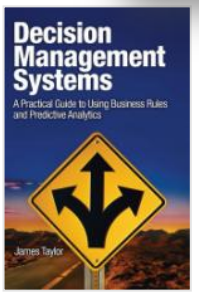
James Taylor, Decision Management Solutions
Bill Franks, Teradata

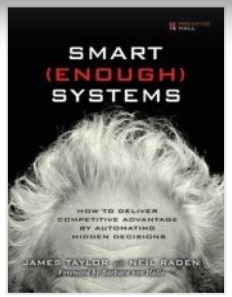
Today's Speakers

James Taylor

CEO

Decision Management Solutions



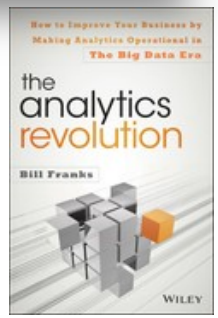


INTERNATIONAL INSTITUTE FOR ANALYTICS

Bill Franks

Chief Analytics Officer

Teradata



INTERNATIONAL INSTITUTE FOR ANALYTICS

Polling question 1

- Polling question 1 in the beginning of the session.
  - What best describes your companies use of R today?
    - No R plans in the near future
    - Exploring or experimenting with R
    - Plans to use R for analytics
    - Actively using R for model development only
    - Actively using R for model development and deployment.

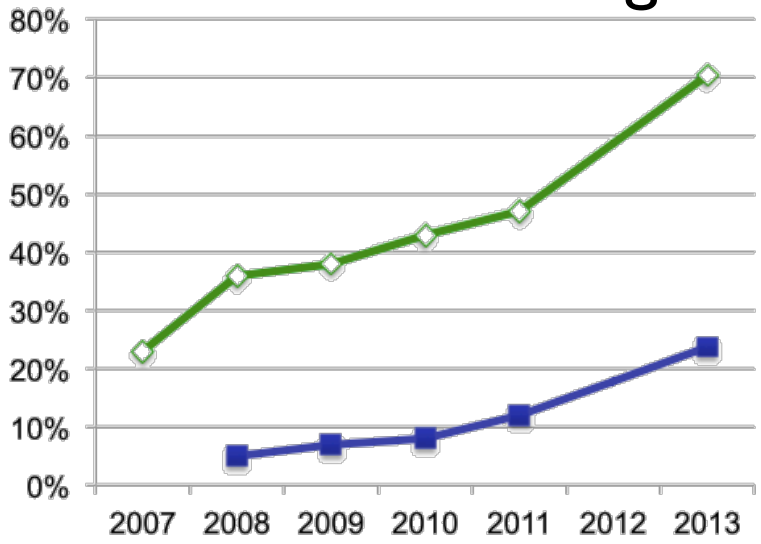
Introducing R

Introducing Open Source R

- The R Project for Statistical Computing
- Interpreted language for statistical computing
  - Extensible
  - Free
  - Open source
  - Since 1997
  - 5,000 Packages

R has become significant in recent years

Data Miners using R

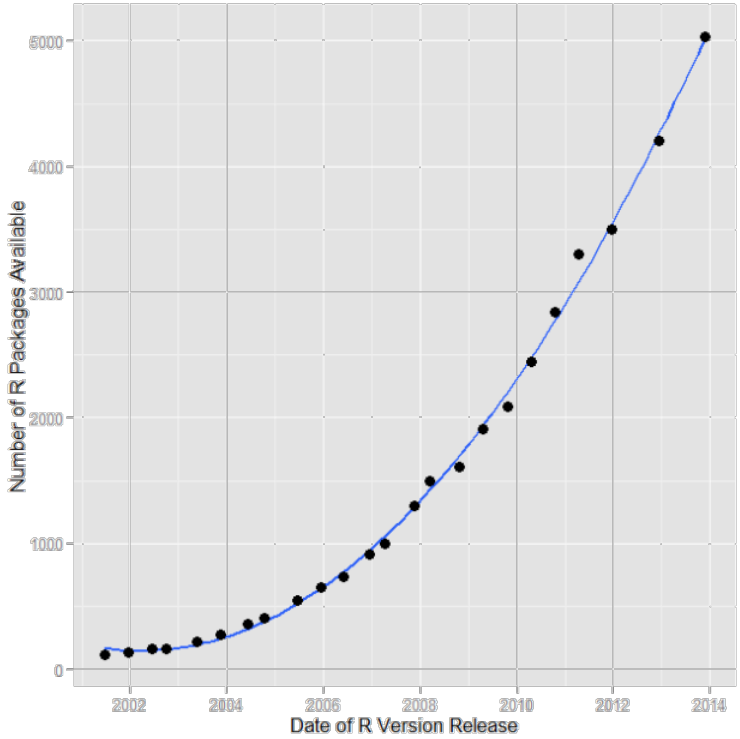


70% of data miners report using R

24% of data miners select R as their primary tool

Rexer Analytics Survey, 2013

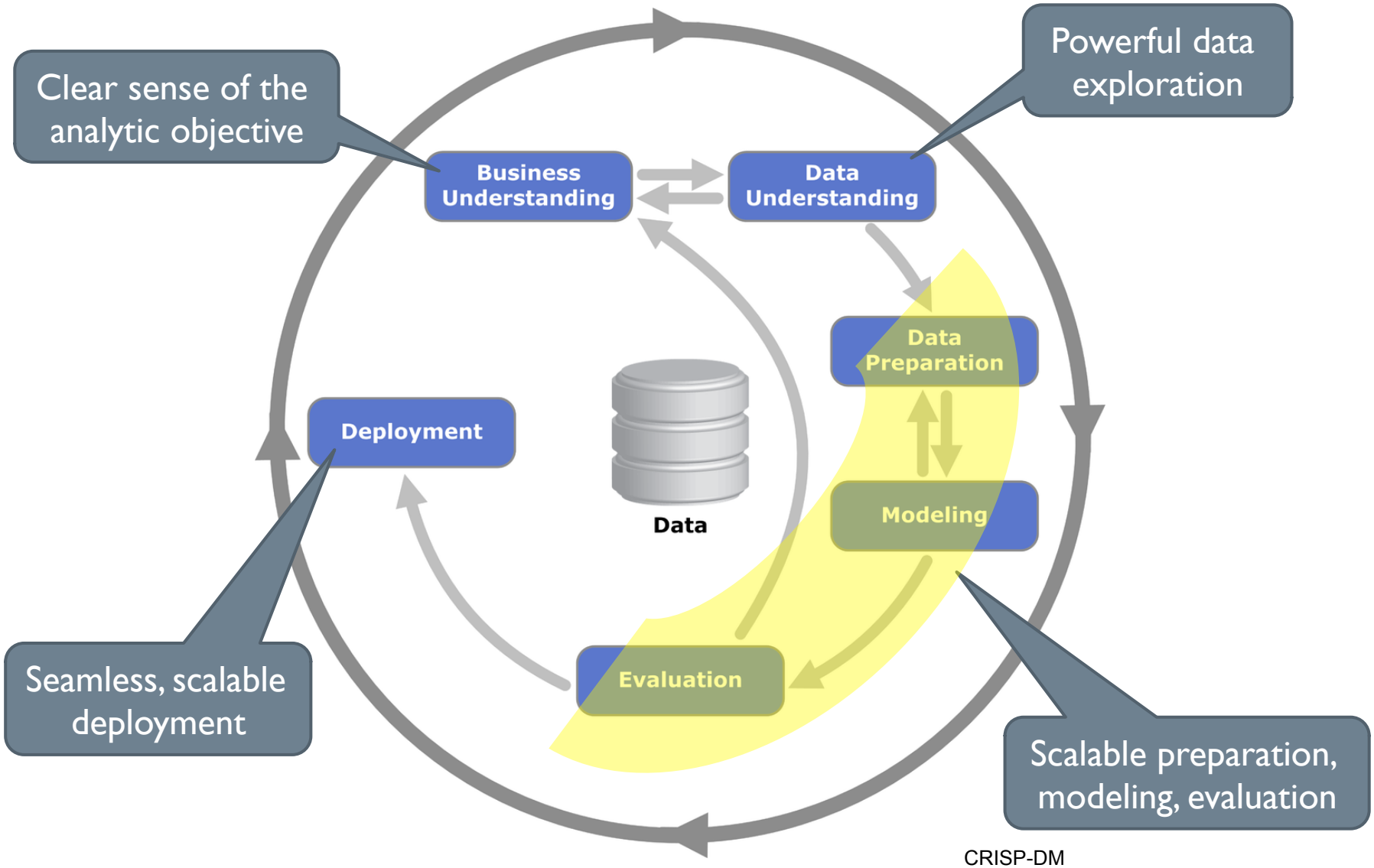
Packages for R



http://r4stats.com/articles/popularity/

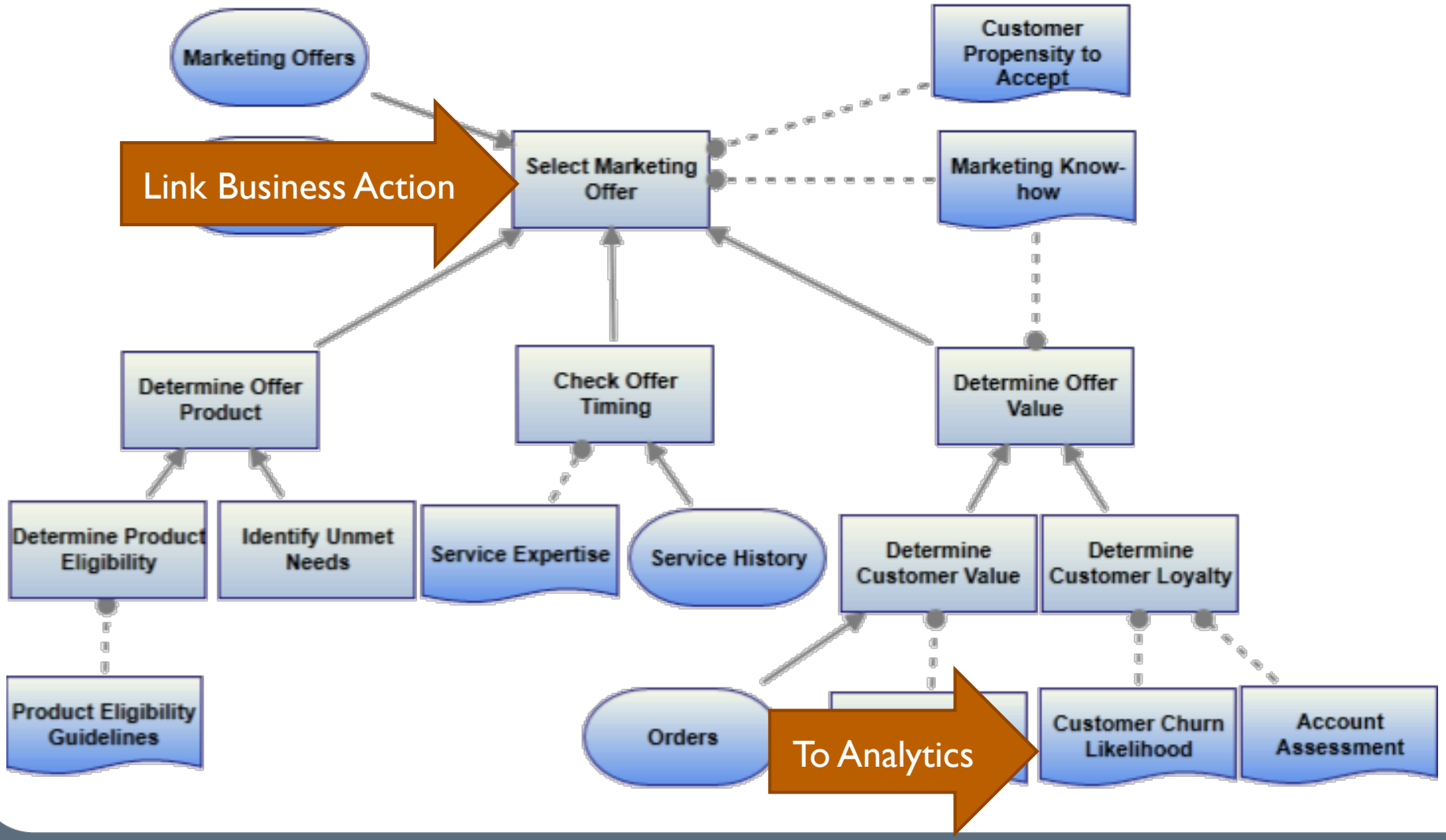
Enterprise Analytic Requirements

Enterprise analytic challenges



CRISP-DM

Clear sense of the analytic objective



Powerful data exploration

Volume

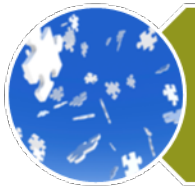
Velocity

Variety

Veracity

Scalable preparation, modeling, evaluation

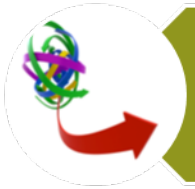
Integrate all the data



Work freely with the data

Model with a wide variety of tools

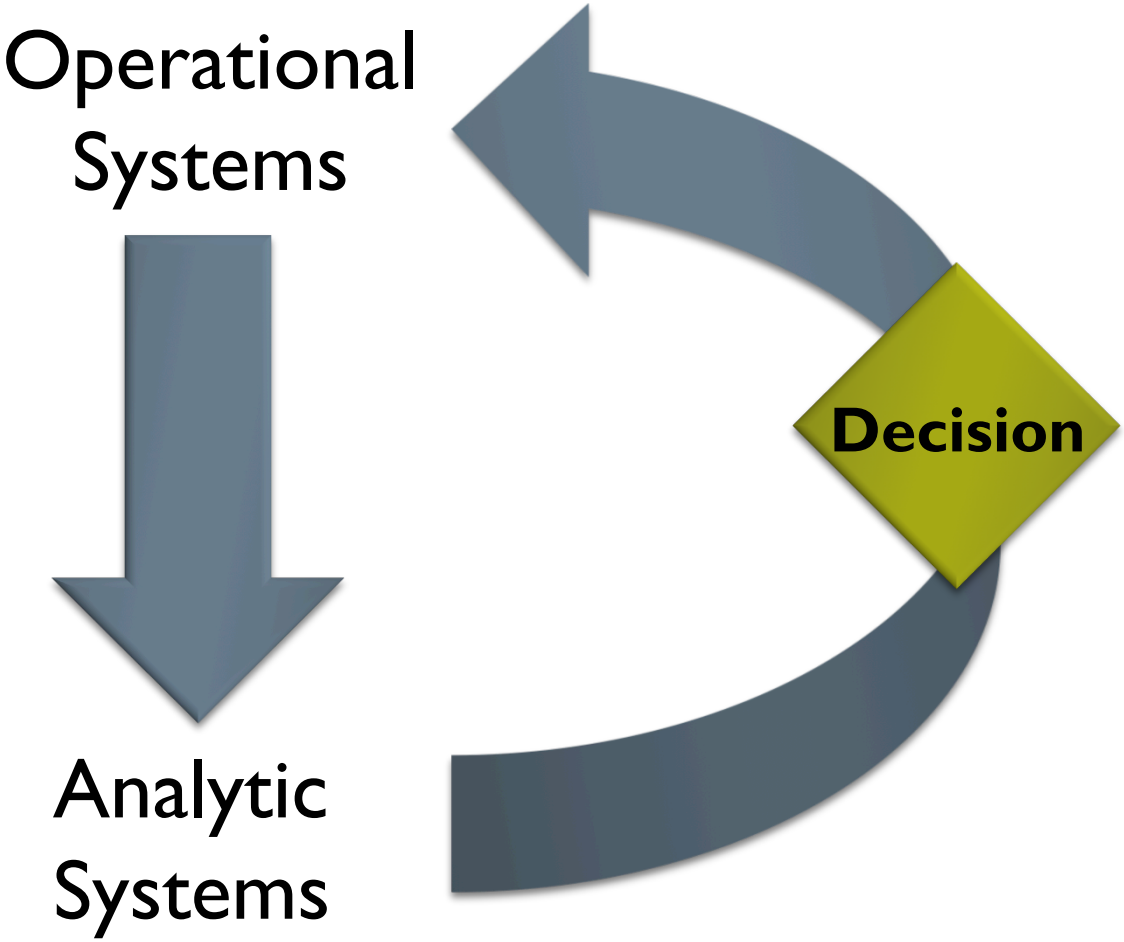
Iterate rapidly to see what works



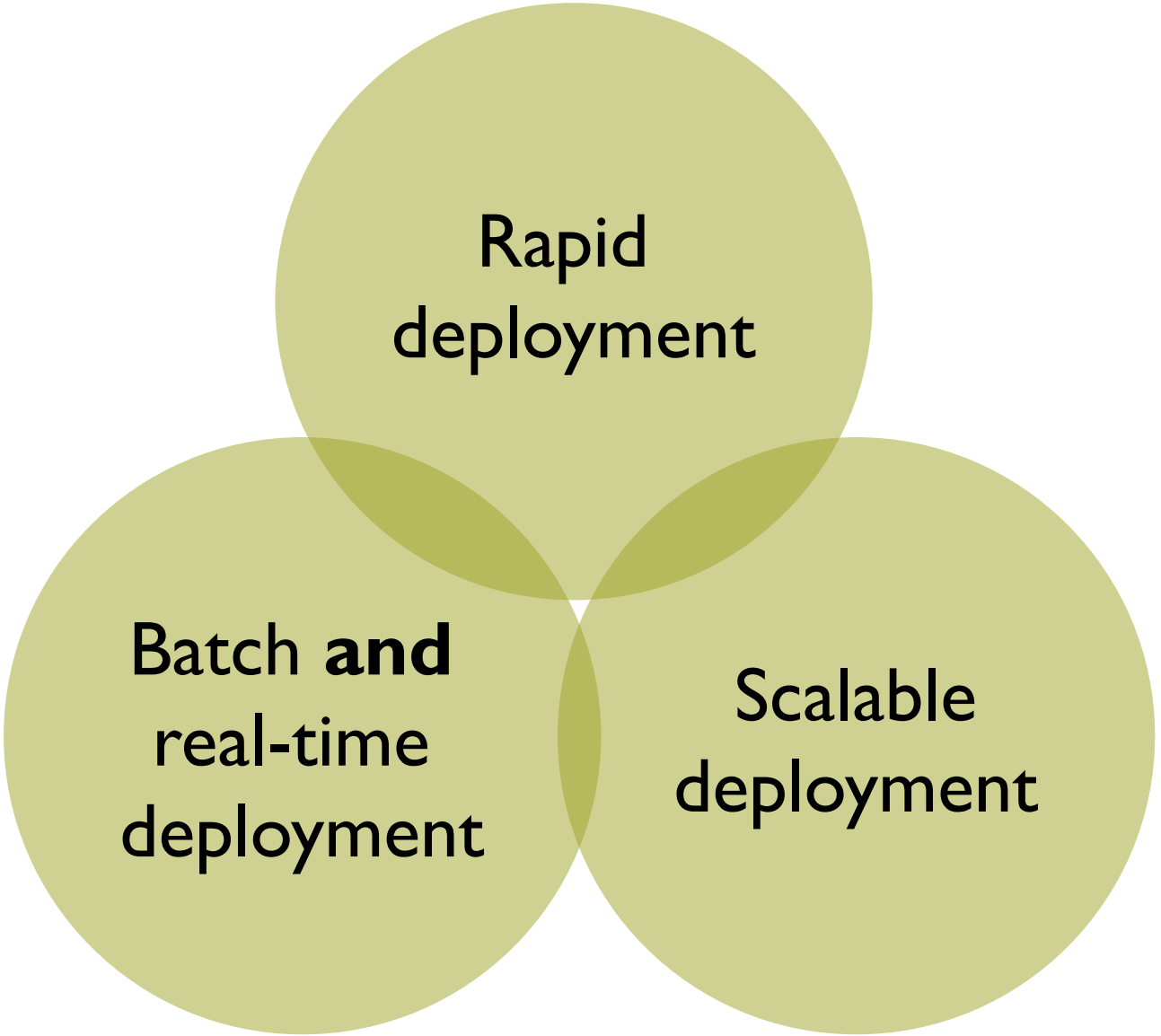
Ensembles matter

Seamless, scalable deployment

Knowing is not enough



Seamless, scalable deployment



Challenges of Open Source R

Enterprise scale analytics and R

- Clear sense of the analytic objective
- Powerful data exploration
- Scalable preparation, modeling, evaluation
- Seamless, scalable deployment
- Complex data integration
- Scaling data understanding
- Time to analyze
- Deployment
- Industrializing for scale

Complex data integration is a challenge

New sources drive data volume in Big Data

Columnar
Hadoop
NoSQL
Relational
Warehouse

Successful analytic teams use increasing variety of data

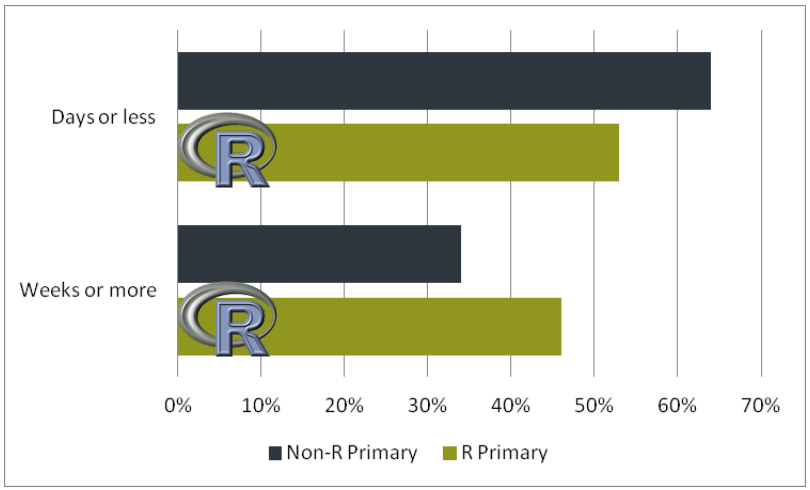
Scaling data understanding is a challenge

More data, more time and effort

Single-threaded
Parallel execution?
In-memory
Forced sampling
Limits iteration

Time to analyze is a challenge

- R users like their tools
  - Lots of algorithms
  - Easy to modify and fine tune
- But
  - Takes longer to do data analysis
  - Tool limit challenges more likely
  - Scaling up a challenge

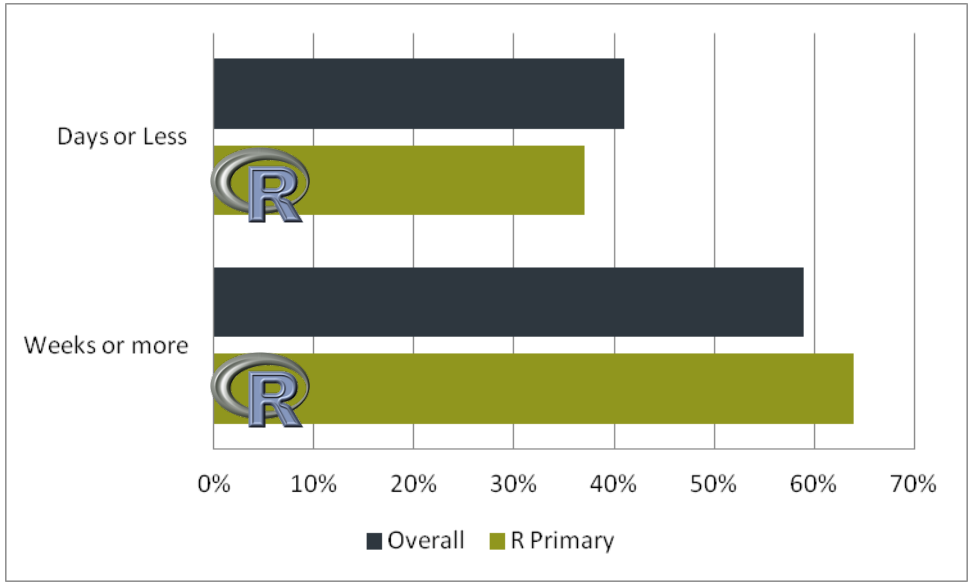


Deployment is a challenge

"Knowing is not enough; we must apply. Willing is not enough; we must do."

Johann Wolfgang von Goethe

1/3 projects have serious deployment challenges



- More likely to not use results
- Unhappy with ease of deployment

Cottage industries don’t scale

Industrialization is a challenge

From

- Local scripts or code
- Hand crafting
- A focus on model **creation**
- Individual creators

To

- Managed workflows
- Automated scale
- A focus on model **management**
- Broad participation and collaboration

R for Enterprise Analytics

- Complex data integration
- Scaling data understanding
- Time to analyze
- Deployment
- Industrializing for scale

Polling question 2:

Polling question 2 between transition from James to Bill.

What are your biggest challenges with R? (select all that apply)

- Complex data integration
- Scaling data understanding
- Time to analyze
- Deployment
- Industrializations
- Others _____

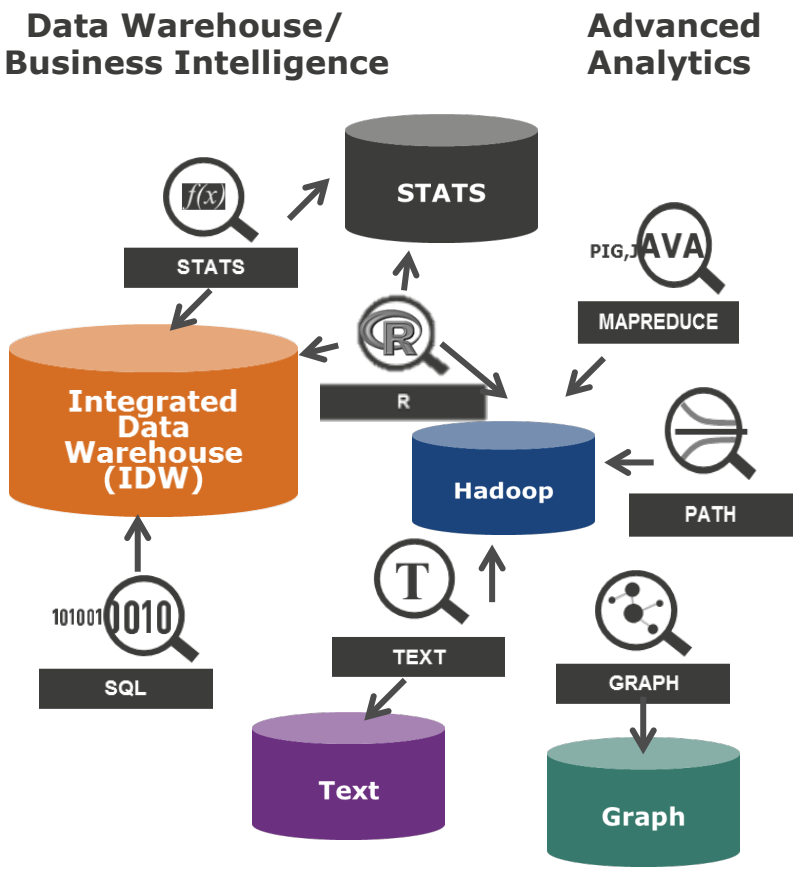
TERADATA

LIFTING THE LIMITATIONS OF OPEN SOURCE R

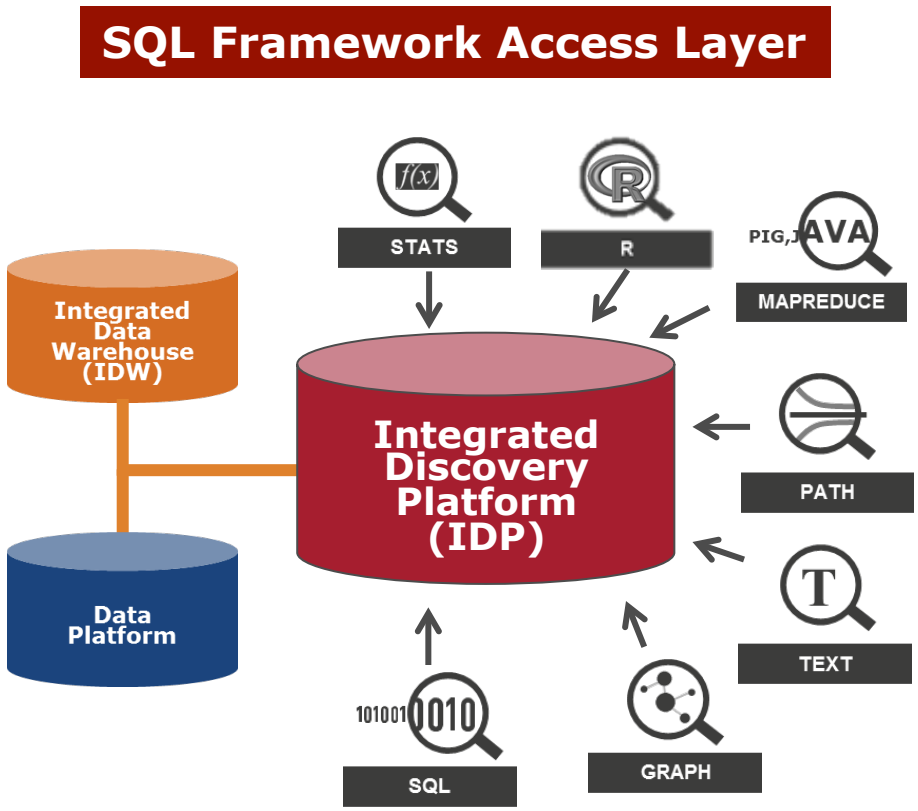
Bill Franks
Chief Analytics Officer
Teradata

The Case For A Discovery Platform

The Problem



The Solution



TERADATA.

Tackle R's Challenges With Aster R!

Complex Data Integration

Scaling Data Understanding

Time To Analyze

Deployment

Industrialization

TERADATA.

Teradata Aster R™

"Making open source R massively scalable & powerful"

Open source R without limitations

- Run open source R across Aster's MPP architecture for high speed parallel processing
- Remove memory and data limitations with in-database processing for massive scalability
- Leverage all data vs. samples for deeper insights

Unmatched ease-of-use and productivity for R users

- Use familiar R client & R language with Aster through Aster R Library
- Expose Aster Discovery Portfolio functions as R functions
- Leverage open source R with no new tools or languages to learn

Powerful analytics combining Aster and R

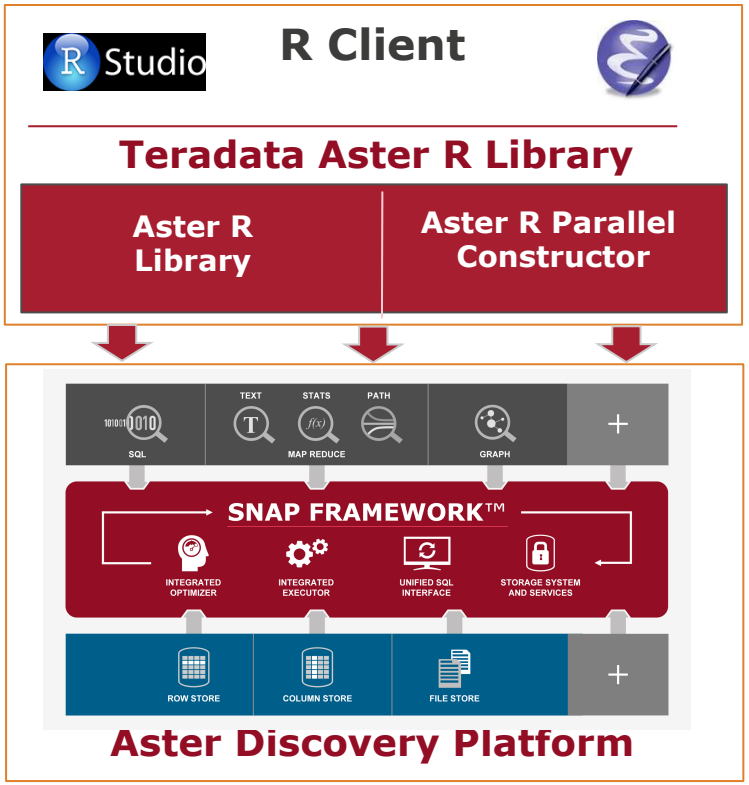
- Combine 100+ Aster Discovery Portfolio functions and 5,000+ R packages for powerful analytics
- Integrate R into Aster's SNAP Framework for rapid discovery
- Empower users with a single comprehensive analytic platform

Teradata Aster R Components

- **Aster R Parallel library**
  - > R functions running in full system level parallel mode
  - > R interface for Aster Discovery Portfolio (MapReduce) functions

- **Aster R Parallel Constructor**
  - > Allows R users to parallelize any R code using split-apply-combine
  - > R engine runs in node independent fashion across all Aster nodes

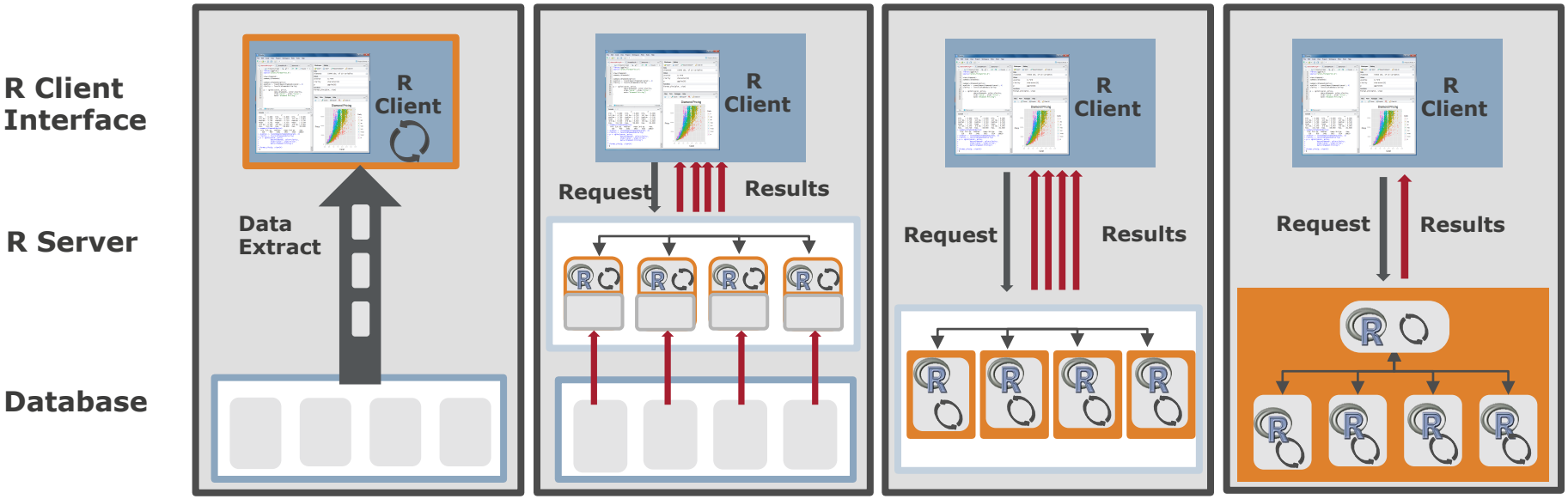
- **R Engine in the SNAP Framework**
  - > Access to any data store, including Hadoop and Teradata
  - > R script can invoke SQL, MapReduce, Graph and R engines
  - > Optimal processing with SNAP's integrated optimizer and executor



TERADATA

R Implementation Options

Where processing takes place



Architecture	Client/Server	Distributed Computing	In-database: Node Level	In-database: System Level
Data access	Data extracted from dbs to client	Data extracted from dbs onto servers	Rows and partitions accessible by the node	Access to all the data within the system
Processing	In-memory of the client	Runs independently in-memory in each server. Coding required for system level parallel processing	Runs independently within each node. Coding required for system level parallel processing	Runs in coordinated parallel manner, often iterating
Results	Single result based on limited data set	Results are returned from each server	Results are returned from each node	Single result is produced from all data

TERADATA

Node Versus System Parallelism

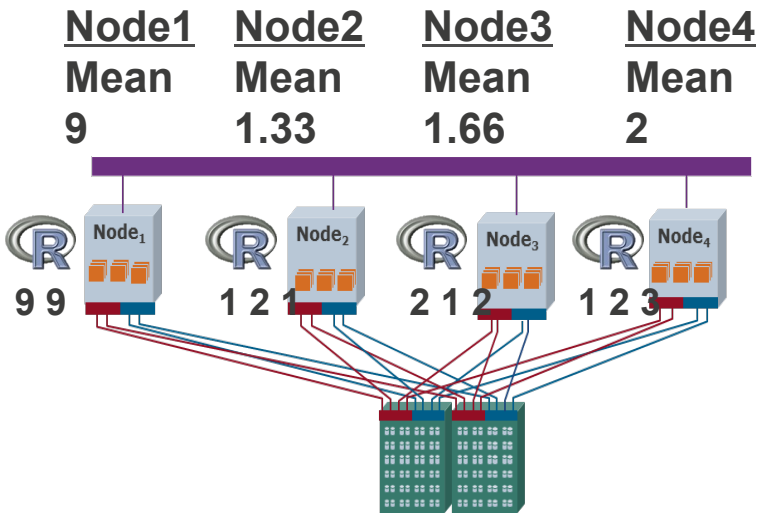
- When R runs independently on each node/server, the onus is on the programmer to code correctly to handle node parallelism

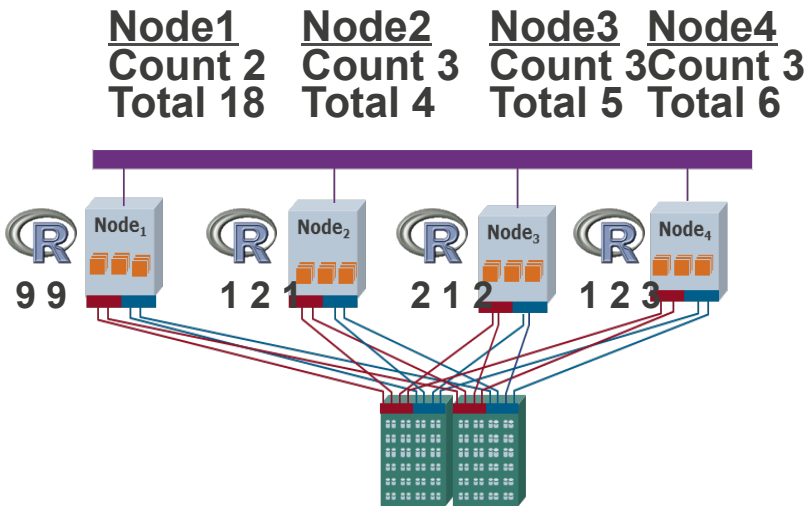
Node Level

1. Find Mean per node
2. Return 1 answer per node or
3. Calculate mean of mean = 3.5 (**X**)

System Level

1. Calculate count and total for each node
2. Aggregate counts (11) and total (33)
3. Calculate mean total/count = 3 (**Correct!**)





TERADATA

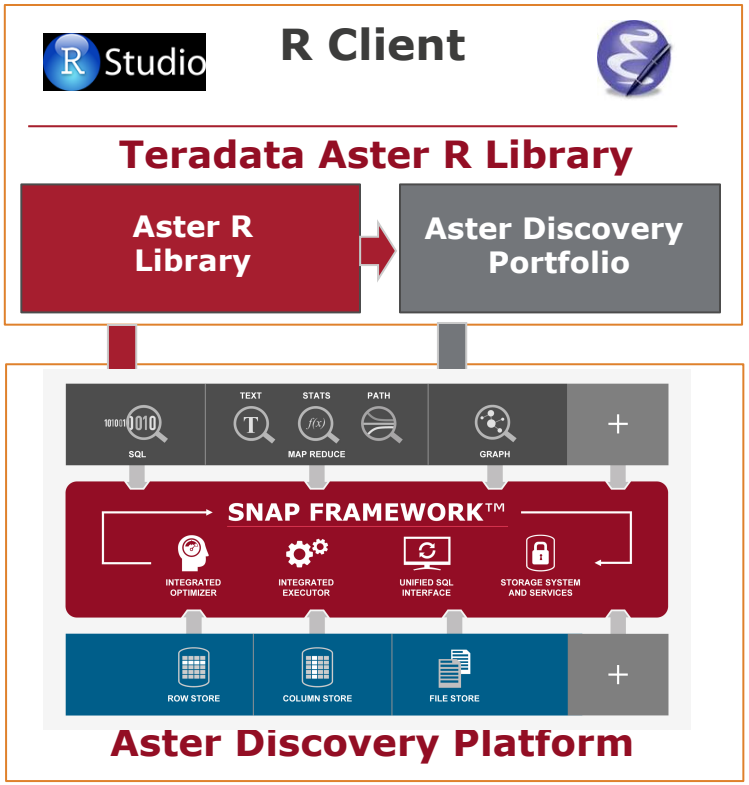
Aster R Library – Prebuilt Parallel Functions

- **Aster R functions run in system parallel mode across all data**
  - > Prebuilt parallel functions to hide the complexity of parallel programming
  - > Allows users to process all data without the need for sampling

- **Familiar R language syntax**

- **Leverages virtual data frames**
  - > Users operate on virtual data frames that point to tables or views in the database

- **Extend R capabilities with Aster Discovery Portfolio functions**
  - > Allows R to invoke Aster MapReduce functions running in system parallel mode



TERADATA

Teradata Aster R Prebuilt Parallel Functions

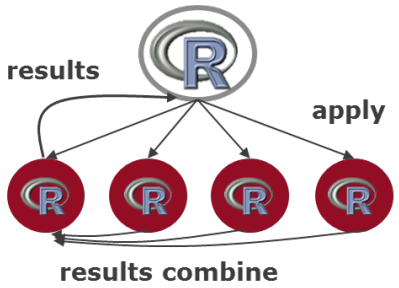
- **Data Access and Movement**
  - > Connect, query, Teradata & Hadoop access via QueryGrid, bulk load & extract, import/export data into tables, read & write csv, and more…
- **Data Management**
  - > Create data frames, refresh, and more…
- **Data Exploration**
  - > Data characteristics, statistics, ranges & distributions, rank, and more…
- **Data Transformation and Manipulation**
  - > Pivot, log parser, unpack/pack, split, matrices, and more…
- **R Operators**
  - > [. [[, $ -, !, +, /, *, %%, ==, !=, and more…
- **Path & Time Series Analysis**
  - > nPath, sessionization, and attribution
- **Statistical Analysis**
  - > Regressions, Naïve Bayes, support vector machine, regressions, correlations, averages, histogram, Principal component analysis, and more…
- **Text Analysis**
  - > Sentiment analysis, text processing, ngram, text classifier, and more..
- **Machine Learning**
  - > Kmeans, basket, collaborative filter, random forest, and more…

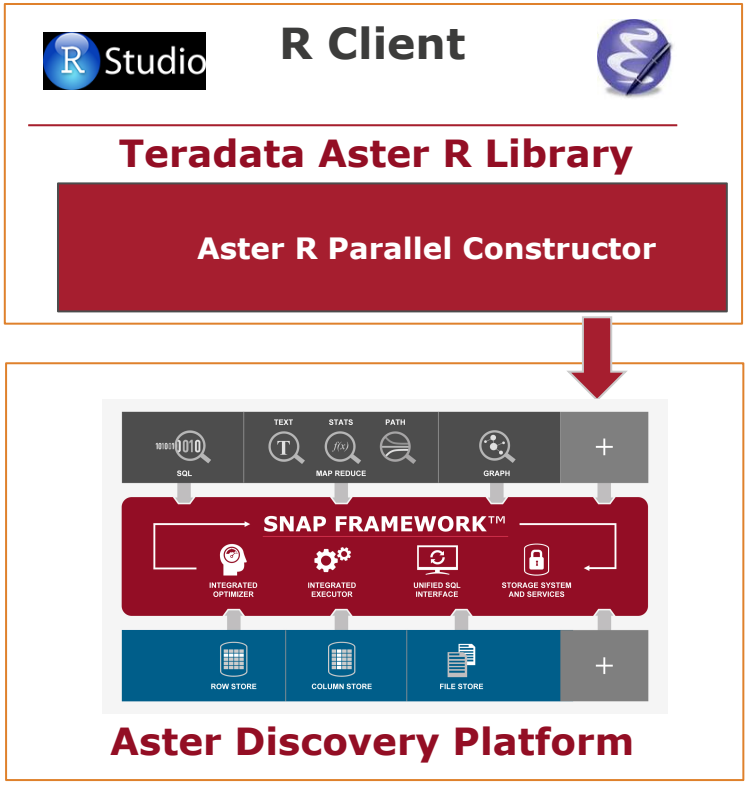
Aster R Parallel Constructor

- **Allows users to run open source R scripts in Aster in parallel**

- **Users can run any open source R code in parallel with ta.apply()**
  - > Accepts a data frame as input
  - > Runs an R script across a single or multiple nodes concurrently
  - > Runs R script using the split-apply-combine strategy
    - Similar to Map-Reduce constructs

- **How It Works**
  - > ta.apply()
    - Apply first R script to each node
    - Apply second R combiner script to the results of the first



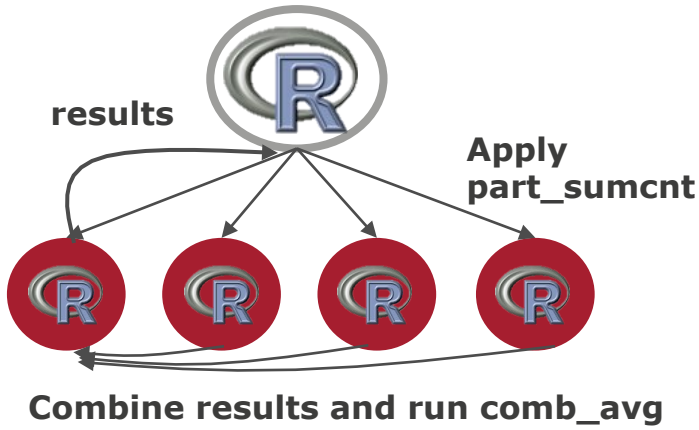


TERADATA

Aster R Parallel Constructor Example

Calculating Average with all the data

- Data is distributed across the nodes
- Ex. Calculate average sales to date




```
banking_df = ta.data.frame("ich_banking")

part_sumcnt <- function(x) list(sum = sum(x), count = length(x))
comb_avg <- function(x) sum(x$sum) / sum(x$count)

ta.aggregateApply(
    banking_df,
    FUN = part_sumcnt,
    FUN.result = "list",
    COMBINER.FUN = comb_avg)
```

Create a view in Aster for parallel data ingest

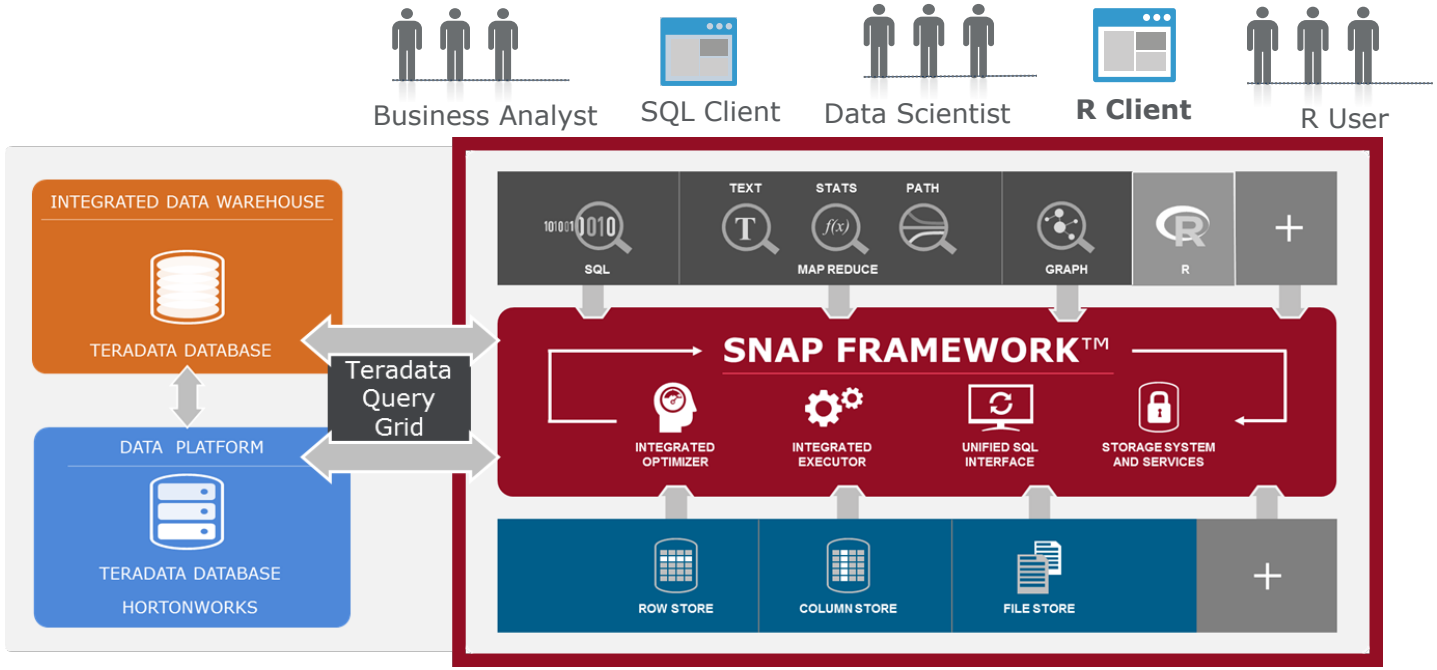
User defined R functions executed in parallel with partitioned data in-place

Runs the function on each vworker node, reduces the results from each partition

TERADATA

Deploying Analytics with Aster R

- **R interface for Aster Discovery Platform**
  - > R users now have access to the powerful Aster Discovery Platform
- **Multi-faceted analytics**
  - > A single program can call SQL, MapReduce, Graph, or R engines
- **Access to any data across the Teradata UDA**
  - > Data from Teradata and Hadoop are accessible through Teradata QueryGrid



TERADATA

TERADATA

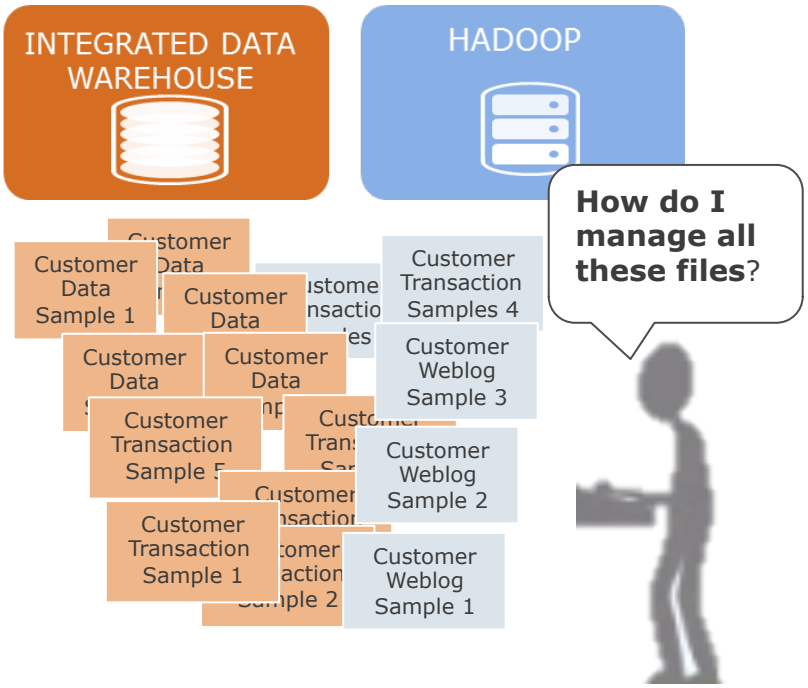
A DAY IN THE LIFE OF AN R ANALYST

Up Your R Game!

Complex Data Integration
Across Multiple Data Sources

- **Traditional R**
  - > Create data frames for big data?

    "Error: cannot allocate vector of size 10 GB"
  - > Forced to sample data



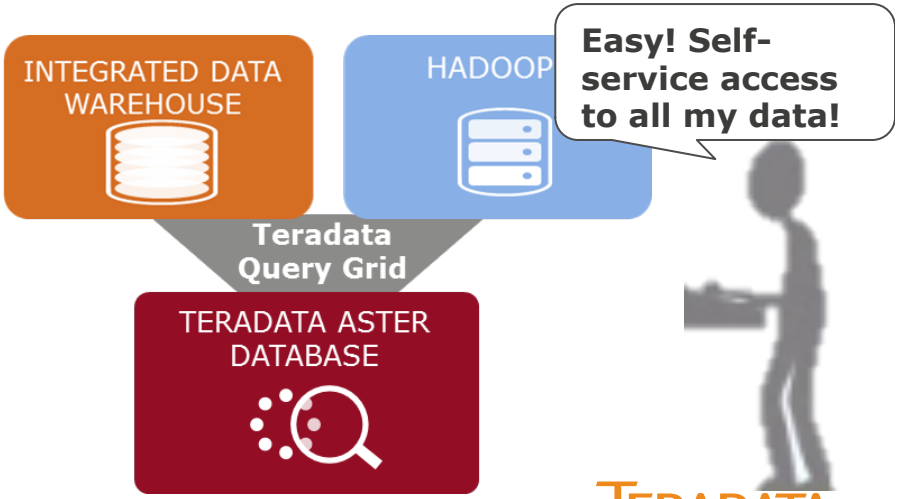
- **Aster R**
  - > Easy Access to Hadoop and DW
    - Create views

```
CREATE VIEW hadoop_view as ( SELECT *
  FROM load_from_hcatalog
(TABLENAME('hcat_table_name')

CREATE VIEW Customer_view as ( SELECT *
  FROM load_from_Teradata
(TABLENAME('customer_table'))
```

    - Create data frames

```
weblog_df <- ta.data.frame("hadoop_view")
customer_df <- ta.data.frame("customer_view")
```



Scaling Data Understanding

- **Traditional R**
  - > Summary function to understand the statistical characteristics of the data
  - > Have a large data set? Memory Error

```
cust1_df <- data.frame (customer_sample 1)
summary (cust1_df);
rm (cust1_df);
```

Create data frame for first customer sample file

Summary function

Free up memory for the next command

  - > Repeat to validate sample
  - > Repeat to understand transaction samples…
  - > Repeat to understand customer web log samples.

WOW! All my data in one command! It's fast too!

- **Teradata Aster R**
  - > df <- ta.data.frame ("customer")
  - > ta.summary (df)

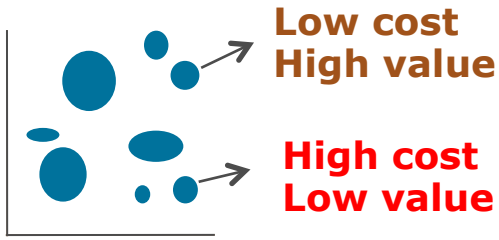
Runs the summary function across all your data!

TERADATA

Time To Analyze

- **Traditional R**
  - > Can't build a cluster on a small sample without losing the most interesting segments
  - > Partitioning data doesn't give me a global view of purchases
    - ca_df <- data.frame ("customer_tbl", where state=CA)
    - kmeans (ca_df)
  - > Writing parallel kmeans will take time…

Build a cluster for customers in California



Now I really understand my customer segments characteristics!

Teradata Aster R:
Complete view of your data

- > global_df <- ta.data.frame("customer tbl")
- > ta.kmeans (global_df, k)

Clusters based on global purchases

Prebuilt parallel kmeans algorithm!

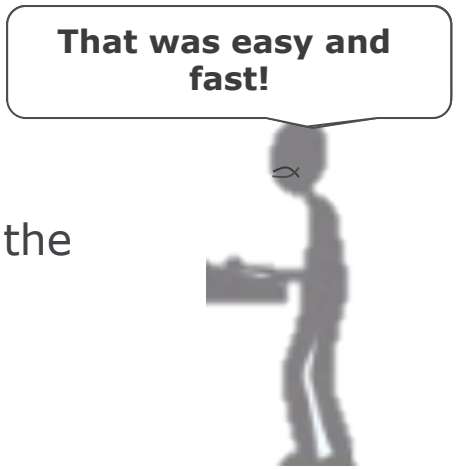
TERADATA

Deployment

- **Traditional R**
  - How do you deploy R models against large volumes of data?
    - Option 1: Get a super computer with LOTS of memory … No budget
    - Option 2: Extract, score, write back into the database … Too labor intensive
    - Option 3: Give to IT to recode my model … Takes time and adds risk
    - Option 4: Use PMML & in-database scoring … Only if IT lets me

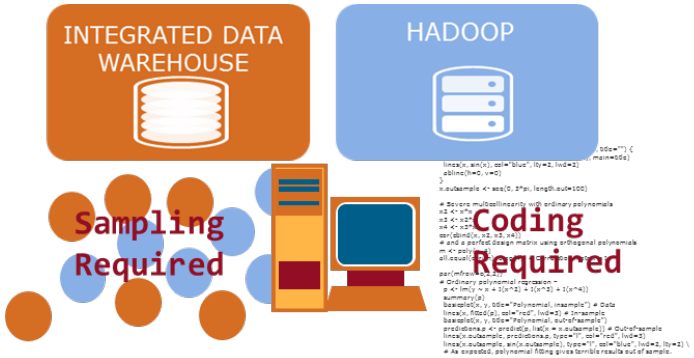
Teradata Aster R:

- Run any open source R package or script concurrently across all nodes or create your own parallel script with the R Parallel Constructors
  - target_segment <- data.frame(customer tbl)
  - ta.apply (target_df, FUN=myRcode)



Industrialization

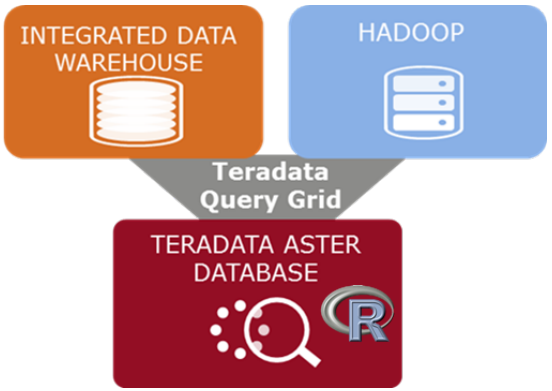
Traditional R



- **Extract Samples** – Inefficient and time consuming
- **Slow Processing** – Single threaded analytics are slow
- **Data Limitations** – Bound by memory…
- **Complex programming** – Parallel programming is hard
- **Deployment challenges** – Often must recode models into SAS or SQL to deploy

Days to Weeks

Aster R



- **Self-serve** - Immediate access to data via Teradata QueryGrid
- **Fast** - Leverage Aster MPP architecture
- **Scalable** – Designed to run against all your data
- **Easy** – No parallel programming required with prebuilt functions
- **Flexible** – Run any open source function in parallel

Seconds to Hours

TERADATA

Teradata Aster R

High Performance Analytic Platform for R with prebuilt parallel analytic functions to process all your data and the flexibility to run any open source R package at scale

To learn more, visit: http://www.teradata.com/Teradata-Aster-R

TERADATA

3rd Polling question

- Polling question 3 at the end of the session.
  - What would you like to learn more about? (select all that apply)
    - How to easily access and integrate data from multiple sources using R
    - How to run R analytics in parallel.
    - How to create models leveraging R, SQL and SQL-MapReduce.
    - Understand model deployment considerations for business analytics
    - Integrating R into production applications.

Thank You

Questions?

James Taylor,
Chief Executive Officer, Decision Management Solutions
james@decisionmanagementsolutions.com
www.decisionmanagementsolutions.com

Bill Franks,
Chief Analytics Officer, Teradata
bill.franks@teradata.com
www.teradata.com