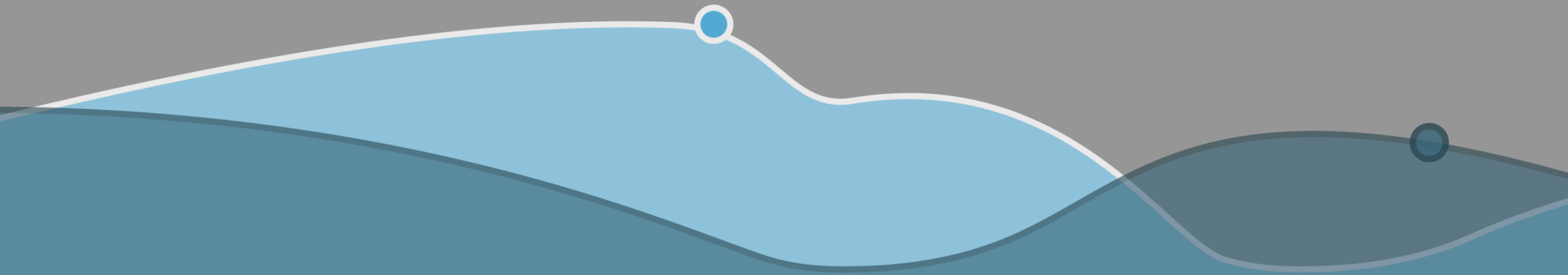
Microsoft

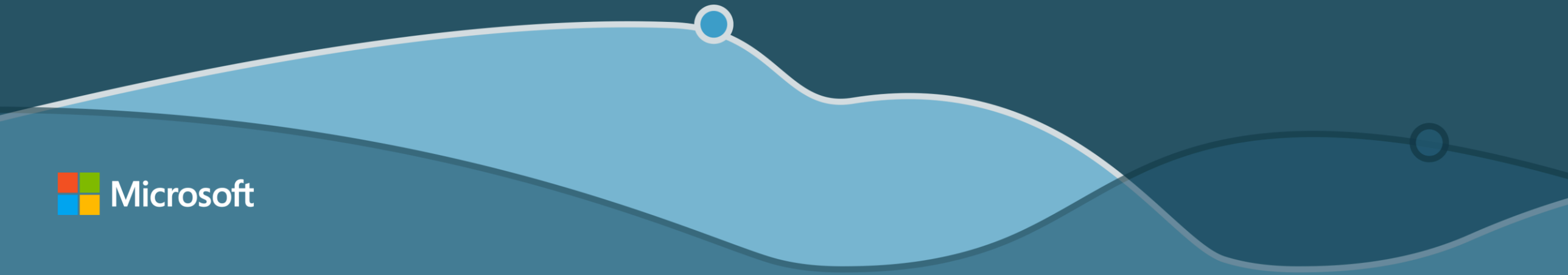
Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC

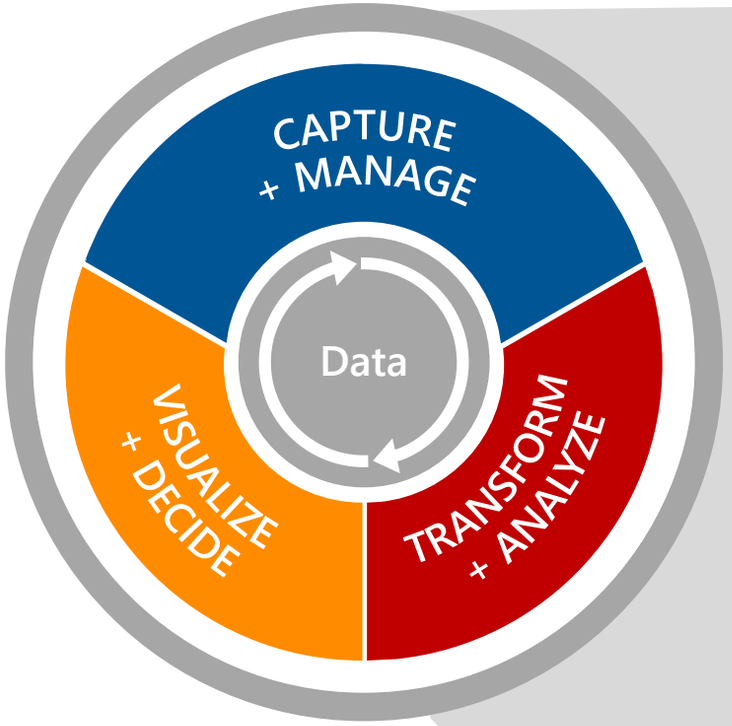


On-Premises Hadoop and Revolution R: Architectures and Solutions

Jamie Olson
Senior Data Scientist, Microsoft



Microsoft



The Microsoft Data Platform

VISUALIZE & DECIDE

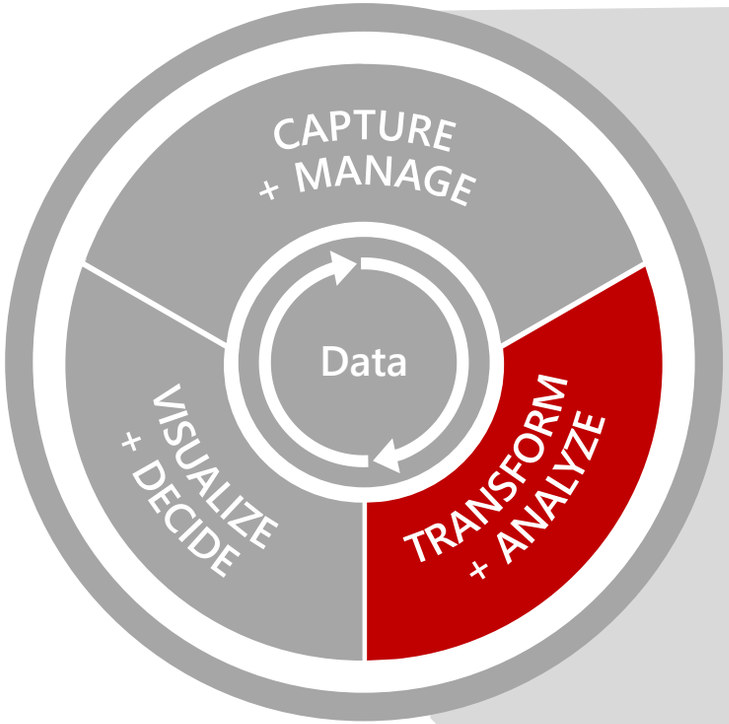
- Applications
- Reports
- Dashboards
- Natural language query
- Mobile

TRANSFORM & ANALYZE

- Orchestration
- Information management
- Complex event processing
- Modeling
- Machine learning

CAPTURE & MANAGE

- Relational
- Non-relational
- NoSQL
- Streaming
- Internal & external



The Microsoft Data Platform

VISUALIZE & DECIDE

- Applications
- Reports
- Dashboards
- Natural language query
- Mobile

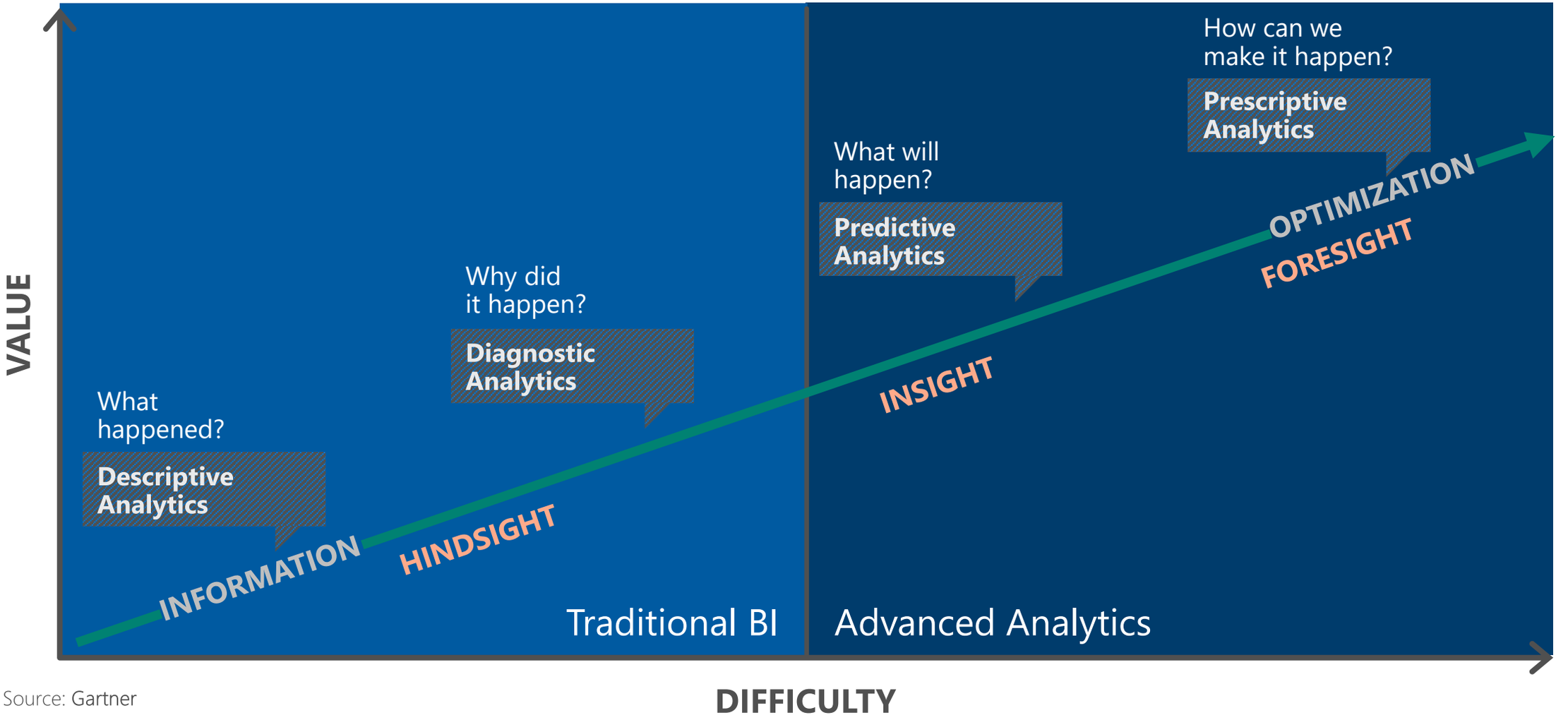
TRANSFORM & ANALYZE

- Orchestration
- Information management
- Complex event processing
- Modeling
- Machine learning

CAPTURE & MANAGE

- Relational
- Non-relational
- NoSQL
- Streaming
- Internal & external

Advanced Analytics
Beyond business intelligence



Source: Gartner

Revolution R Enterprise

R: What Is It?

Download the White Paper
R is Hot
bit.ly/r-is-hot

- A Language Platform…
  - A Procedural Language optimized for Statistics and Data Science
  - A Data Visualization Framework
  - Provided as Open Source
- A Community…
  - 2.5M+ Statistical Analysis and Machine Learning Users
    - Taught in Most University Statistics Programs
    - Active User Groups Across the World
- An Ecosystem
  - CRAN: 6000+ Freely Available Algorithms, Test Data and Evaluations
  - Many Applicable to Big Data If Scaled

CRAN: Resources For All Fields of Analysis

CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know have no idea they exist. As an effort to make them more widely known I thought I'd jazz up the index page. Images are free to use, and got from SXC stock photo site. Visual puns are mine. Task View links go to the cran.r-project.org site and not a mirror.

Bayesian Inference

Applied researchers interested in Bayesian statistics are increasingly attracted to R because of the ease of which one can code algorithms to sample...[more]

Natural Language Processing

This CRAN task view contains a list of packages useful for natural language processing....[more]

Analysis of Spatial Data

Base R includes many functions that can be used for reading, vizualising, and analysing spatial data. The focus in this view is on "geographical" spatial...[more]

Chemometrics and Computational Physics

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of...[more]

Analysis of Pharmacokinetic Data

The primary goal of pharmacokinetic (PK) data analysis is to determine the relationship between the dosing regimen and the body's exposure to the drug as...[more]

Clinical Trial Design, Monitoring, and Analysis

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including...[more]

Official Statistics & Survey Methodology

This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide...[more]

Survival Analysis

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an...[more]

Cluster Analysis & Finite Mixture Models

This CRAN Task View contains a list of packages that can be used for finding groups in data and modelling unobserved cross-sectional heterogeneity. Many...[more]

Phylogenetics, Especially Comparative Methods

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical...[more]

Time Series Analysis

Base R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are...[more]

Probability Distributions

For most of the classical distributions, base R provides probability distribution functions (p), density functions (d), quantile functions (q), and...[more]

Multivariate Statistics

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this...[more]

Robust Statistical Methods

Robust (or "resistant") methods for statistics modelling have been available in S from the start, in R in package stats (e.g., median(), mean(*, trim = .),...[more]

Computational Econometrics

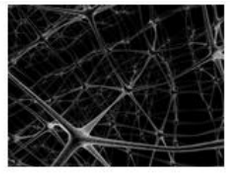
Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many...[more]

Optimization and Mathematical Programming

This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics...[more]

Analysis of Ecological and Environmental Data

This Task View contains information about using R to analyse ecological and environmental data....[more]



Machine Learning & Statistical Learning

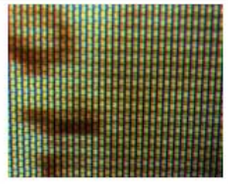
Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually...[more]

Statistics for the Social Sciences

Social scientists use a wide range of statistical methods. To make the burden carried by this task view lighter, I have suppressed detail in some areas that...[more]

Design of Experiments (DoE) & Analysis of Experimental Data

This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements,...[more]



Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic...[more]

gRaphical Models in R

Wikipedia defines a graphical model as a graph that represents independencies among random variables by a graph in which each node is a random variable, and...[more]

Empirical Finance

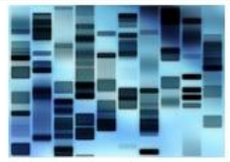
This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic....[more]

High-Performance and Parallel Computing with R

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are...[more]

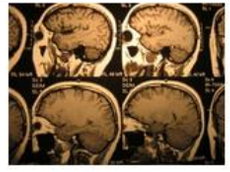
Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better...[more]



Statistical Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs)...[more]



Medical Image Analysis

This task view is for input, output, and analysis of medical imaging files....[more]

Psychometric Models and Methods

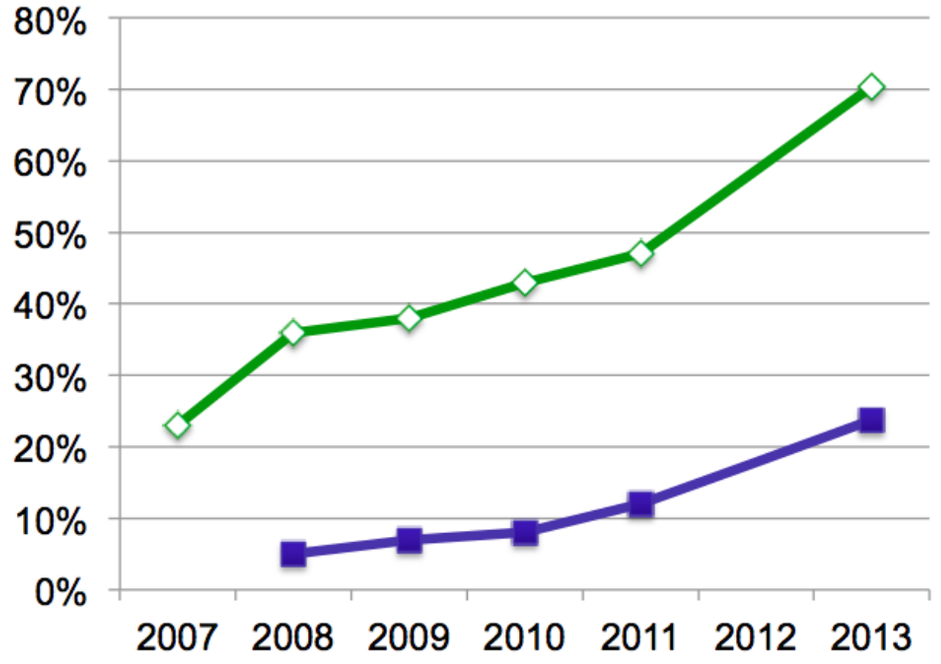
Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked...[more]

R's popularity is growing rapidly

More at blog.revolutionanalytics.com/popularity

R Usage Growth

Rexer Data Miner Survey, 2007-2013



- Rexer Data Miner Survey

Language Popularity

IEEE Spectrum Top Programming Languages

Language Rank	Types	Spectrum Ranking
1. Java		100.0
2. C		99.2
3. C++		95.5
4. Python		93.4
5. C#		92.2
6. PHP		84.6
7. Javascript		84.3
8. Ruby		78.6
9. R		74.0
10. MATLAB		72.6

#9: R

- IEEE Spectrum, July 2014

R Limitations

- Data Flows Promise to Overwhelm Open Source R
  - In-Memory Data Access Model
  - Lack of **implicitly** Parallel Computation
  - Requires Data Movement Prior to Analysis
- As Enterprise Dependence Rises
  - Inadequacy of Community Support
  - Lack of Guaranteed Support Timeliness

Revolution Analytics at a Glance

Who We Are

Only provider of commercial big data big analytics platform based on open source R statistical computing language

Our Software Delivers

Scalable Performance: Distributed & parallelized analytics
Cross Platform: Write once, deploy anywhere
Productivity: Easily build & deploy with latest modern analytics

Our Services Deliver

Knowledge: Our experts enable you to be experts
Time-to-Value: Our Quickstart program gives you a jumpstart
Guidance: Our customer support team is here to help you

Customers

300+ Global 2000

Global Presence

North America / EMEA / APAC

Global Industries Served

Financial Services
Digital Media
Government
Health & Life Sciences
High Tech
Manufacturing
Retail
Telco

Revolution Analytics

Our Vision:

- R becomes the de-facto standard for enterprise predictive analytics

Our Mission:

- Drive enterprise adoption of R by providing enhanced R products tailored to meet enterprise challenges

Revolution Analytics Builds & Delivers:

Software Products

- Stable Distributions
- Broad Platform Support
- Big Data Analytics
- Application Integration
- Deployment Integration
- Agile Development Tooling
- Future Platform Support

Support & Services

- Commercial Support Programs
- Training Programs
- Professional Services

Community Programs

- Academic Support Programs
- Contributions to Open Source R
- Open Source Extensions
- Sponsorship of R User Groups

The Revolution R Product Suite

Revolution R Open

- Free and open source R distribution
- Enhanced and distributed by Revolution Analytics

Revolution R Enterprise

- Secure, Scalable and Supported Distribution of R
- With proprietary components created by Revolution Analytics

Key Value from Revolution R Enterprise

- Enhanced Open Source Delivers:
  - **Simplicity**
    - R Skills Transfer / Lower cost of Talent
    - Ease of Integration with Other Analytics Packages & Data
    - Access to Huge Libraries of R Analytical Algorithms
  - **Speed**
    - Intel-Optimized Computation
  - **Capability**
    - 6500+ Algorithm & Connector Packages Available for Free in CRAN

- Revolution R Enterprise Delivers:
  - **Speed**
    - Distributed Computation using Parallelized Algorithms
    - In-Hadoop & In Teradata Analysis
  - **Scale**
    - No In-Memory Limitations
    - Efficient Data Storage Formats
  - **Stability**
    - Commercial Support & Services
    - Platform Longevity
  - **Time-to-Results**
    - Powerful IDE & Strong Integration
    - Multi-Platform Scoring
  - **Compatibility**
    - Web Services Based Integration Platform

ScaleR Functions & Algorithms

Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

Subsample (observations & variables)

- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression

Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

Cluster Analysis

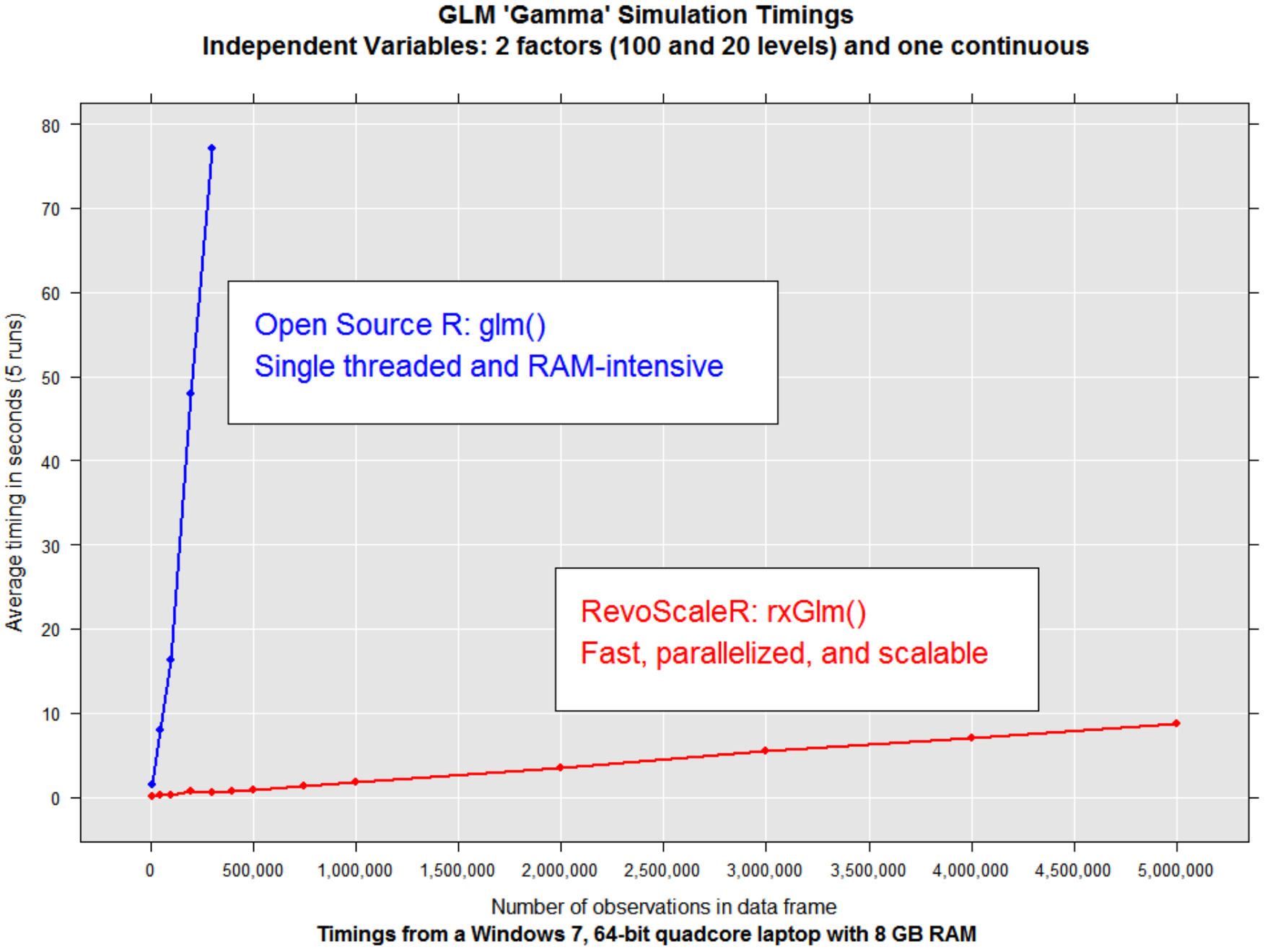
- K-Means

Classification

- Decision Trees
- Decision Forests
- **Gradient Boosted Decision Trees**
- **Naïve Bayes**

Combination

- **PEMA-R API**
- rxDataStep
- rxExec


GLM 'Gamma' Simulation Timings
Independent Variables: 2 factors (100 and 20 levels) and one continuous
Average timing in seconds (5 runs)
0
10
20
30
40
50
60
70
80
Open Source R: glm()
Single threaded and RAM-intensive
RevoScaleR: rxGlm()
Fast, parallelized, and scalable
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
3,500,000
4,000,000
4,500,000
5,000,000
Number of observations in data frame
Timings from a Windows 7, 64-bit quadcore laptop with 8 GB RAM

ScaleR is 10's to 100's of Time Faster Than Open Source

No RAM Limits

- Open Source R Exceeds RAM and Fails
- RRE Scales Linearly Well Beyond RAM Limits

Faster Algorithms

- As data grows ScaleR optimization becomes apparent

File Name	Compressed File Size (MB)	No. Rows	Open Source R (secs)	Revolution R (secs)
Tiny	0.3	1,235	0.00	0.05
V. Small	0.4	12,353	0.21	0.05
Small	1.3	123,534	0.03	0.03
Medium	10.7	1,235,349	1.94	0.08
Large	104.5	12,353,496	60.69	0.42
Big (full)	12,960.0	123,534,969	Memory!	4.89
V.Big	25,919.7	247,069,938	Memory!	9.49
Huge	51,840.2	494,139,876	Memory!	18.92

- Public US Flight Data
- Linear Regression on Arrival Delay
- Run on 4 core laptop, 16GB RAM and 500GB SSD

Scalable Performance: 1 Billion Rows Logit

3.7x Faster than SAS HPA

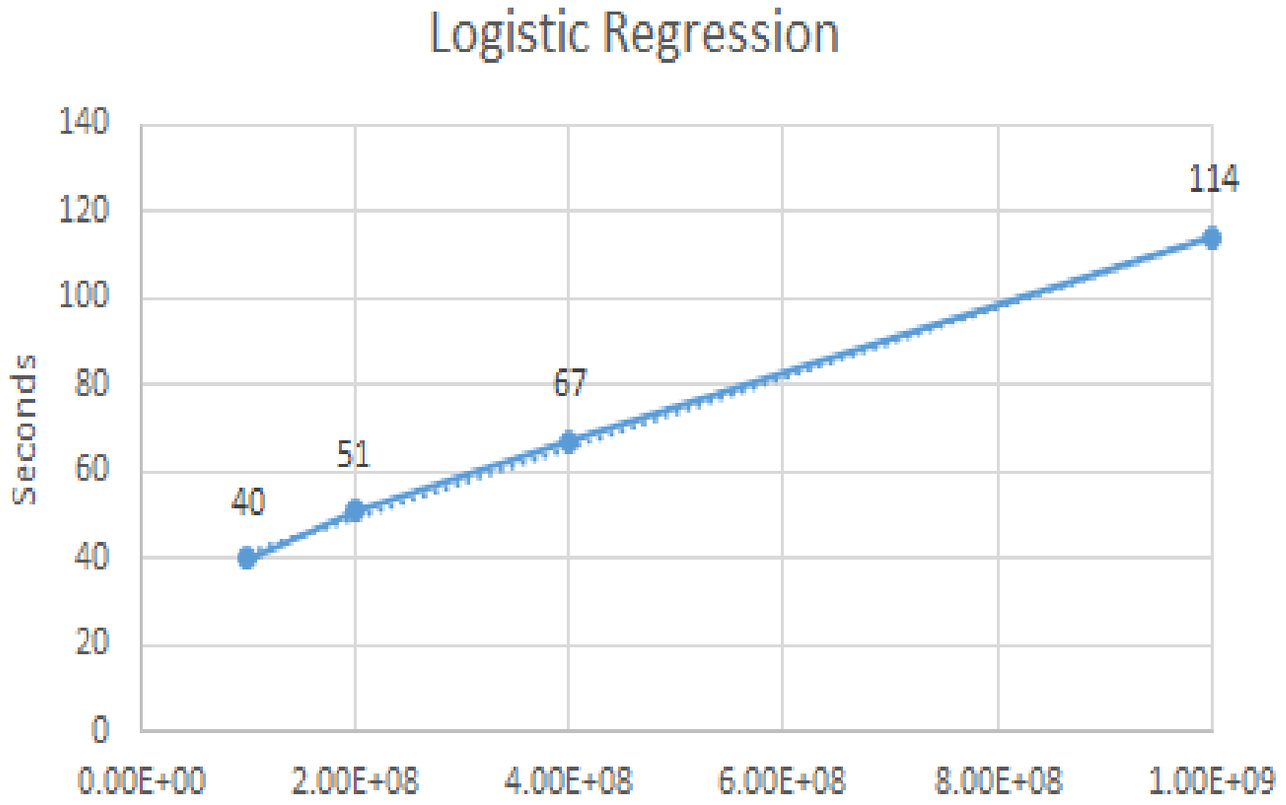
SAS on 384 Core Greenplum: 80 sec.

RRE on 72 Core EDW: 114 sec.

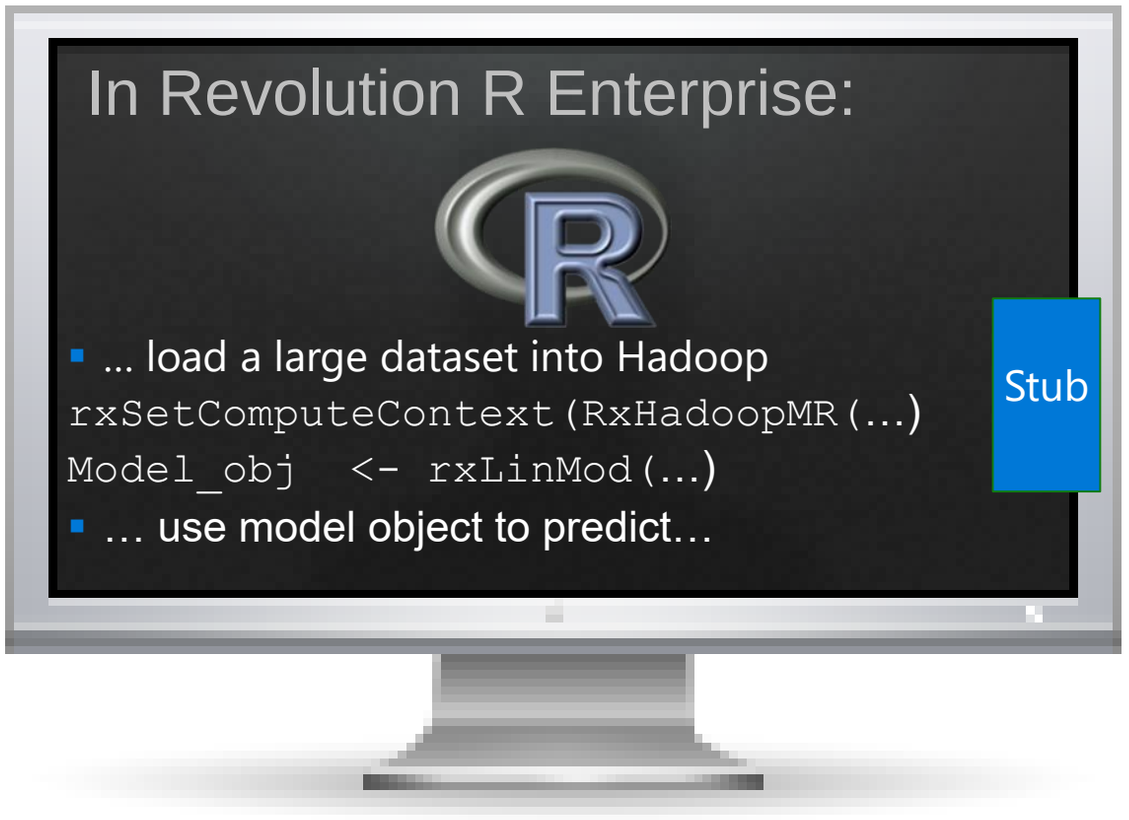
Scales linearly with cores

6 node were 3 x faster than 2 nodes

Scales Linearly with Volume (chart)

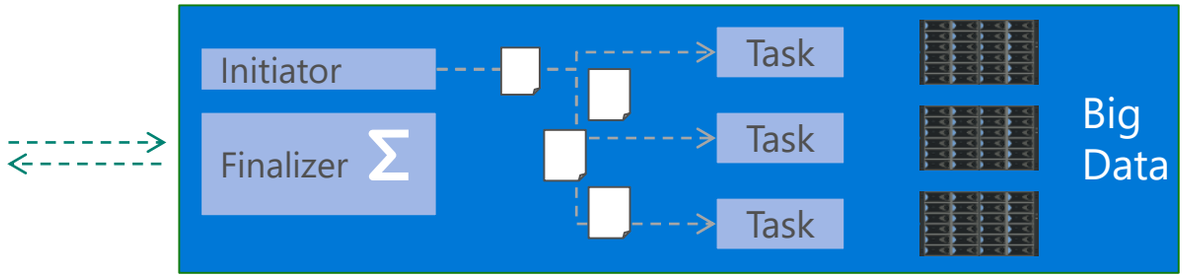


Easy, Portable, Parallelized Analytics



Remote Predictive Algorithms

1. Starts A Master Process
2. Distribute Work
   1. Threading? Cores? Sockets? Nodes?
   2. Available RAM?
   3. Location of Data?
3. Master Tasks for Each Node



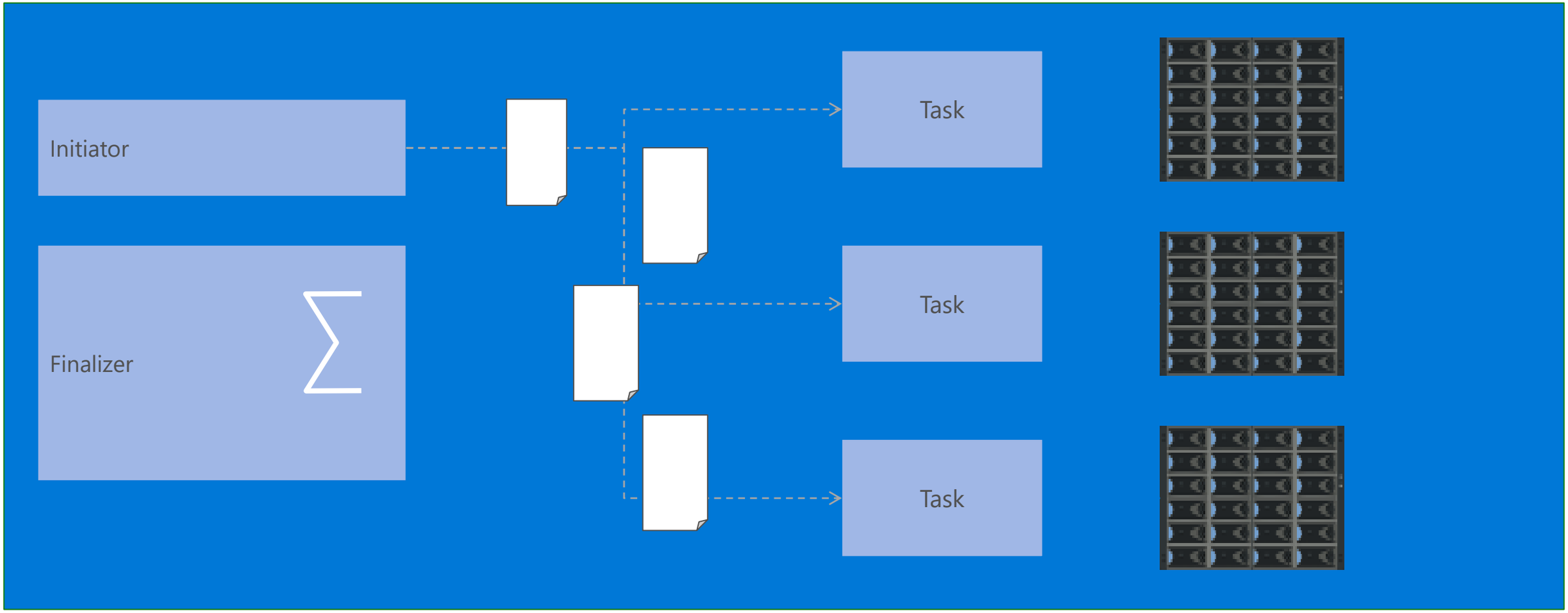
Move Logic To The Data

4. Master Initiates Distributed Work
   1. Hadoop Schedules Mapper for Each Split
   2. Algorithm Computes Intermediate Result
   3. Reducer Combines Intermediate Results
5. Master Process Evaluates Completion
6. Returns Consolidated Answer to Script

Easy Analytics

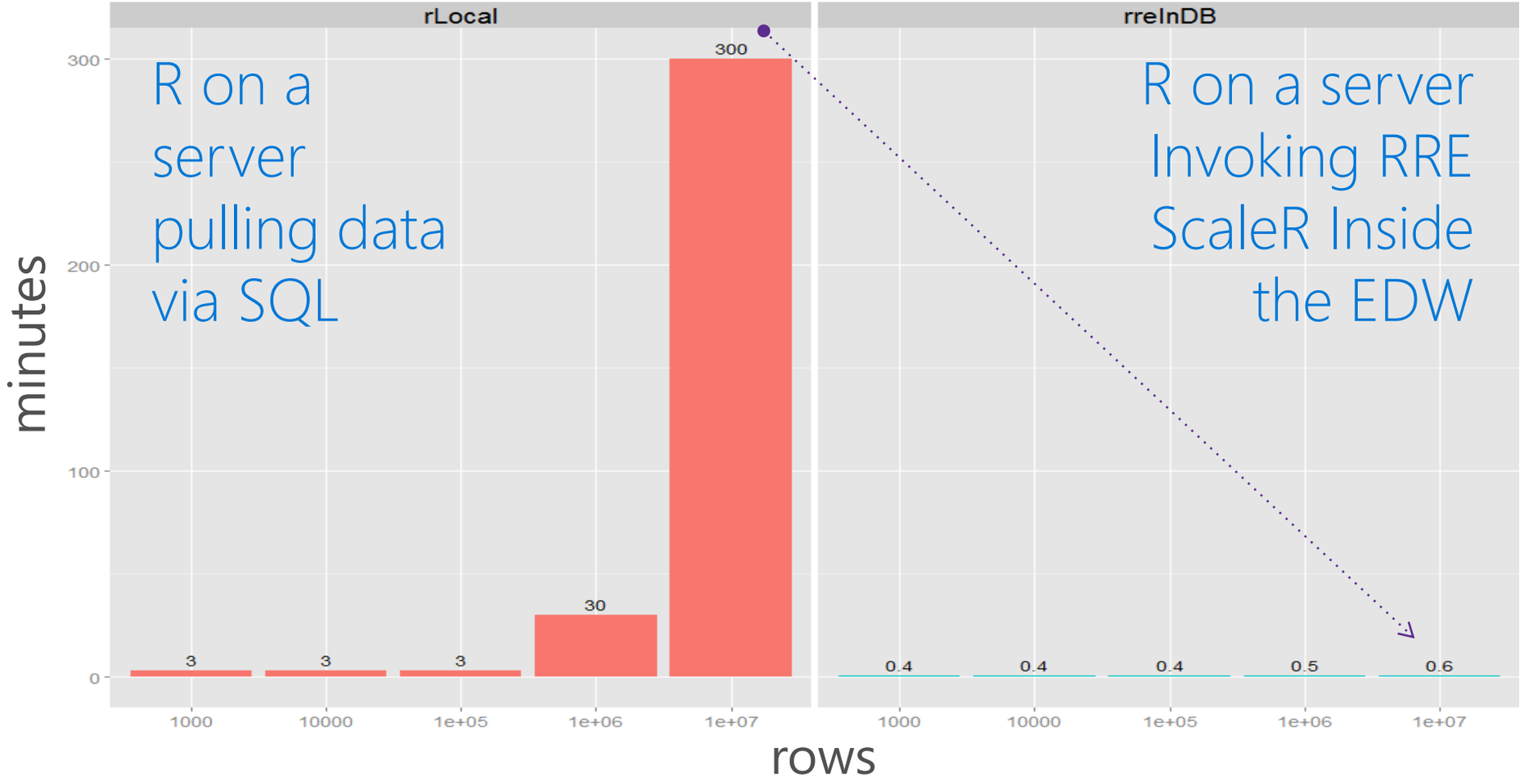
```
myHadoopCluster <- RxHadoopMR(consoleOutput=TRUE)
rxSetComputeContext(myHadoopCluster)
hdfsFS <- RxHdfsFileSystem()
model <-
    rxSummary(ArrDelay~CRSDepTime+DayOfWeek,
        data = airDS)
model
```

Portable and Parallelized Analytics



In-Database Acceleration

5+ hours to 40 seconds: Recommendation is that this now become the defacto productionalization process



Example Performance Comparison to SAS

October 25, 2012

Allstate compares SAS, Hadoop and R for Big-Data Insurance Models

At the Strata conference in New York today, Steve Yun (Principal Predictive Modeler at Allstate's Research and Planning Center) described the various ways he tackled the problem of fitting a generalized linear model to 150M records of insurance data. He evaluated several approaches:

1. Proc GENMOD in SAS
2. Installing a Hadoop cluster
3. Using open-source R (both on the full data set, and on using sampling)
4. Running the data through Revolution R Enterprise

Steve described each of the approaches as follows.

Generalised Linear Model

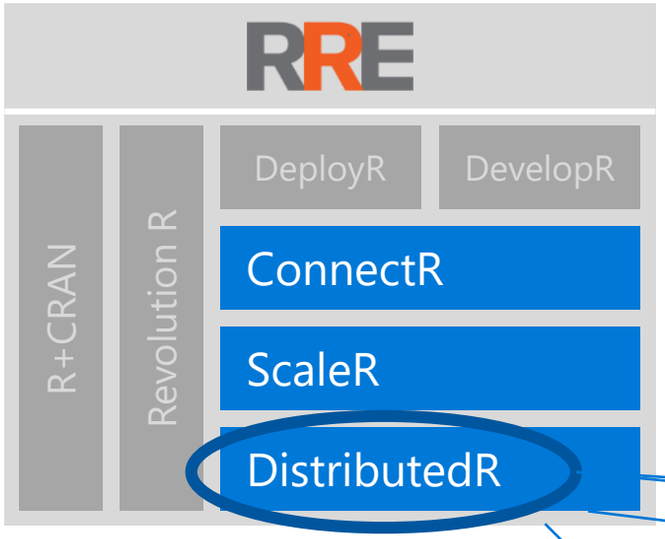
- 150 million observations
- 70 degrees of freedom

Software	Platform	Time to fit
SAS	16-core Sun Server	5 hours
rmr / map-reduce	10-node (8 cores / node) Hadoop cluster	> 10 hours
Open source R	250-GB Server	Impossible (> 3 days)
RevoScaleR	5-node (4 cores / node) LSF cluster	5.7 minutes

So what have we learned:

- SAS works, but is slow.
- The data is too big for open-source R, even on a very large server.
- Hadoop is not a right fit
- Revolution R Enterprise gets the same results as SAS, but about 50x faster.

DistributedR



Delivers High Performance Parallel Distributed Analytics Across Individual and Clustered Systems

Hadoop

- Cloudera
- Hortonworks
- MapR
- Apache Spark (coming soon)

MPI

- IBM Platform LSF
- Microsoft HPC Clusters

Local Files

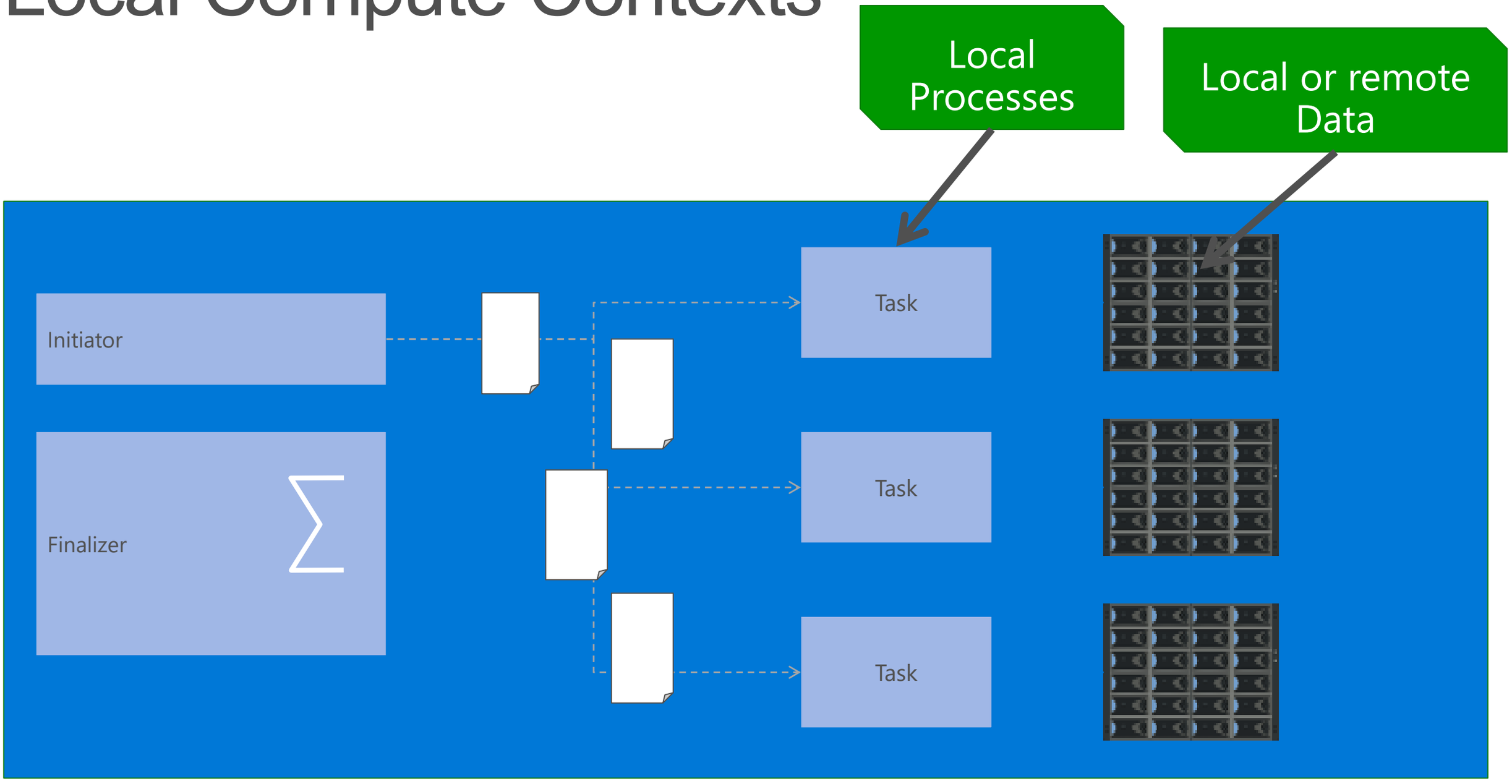
- Red Hat
- SuSE Servers
- Windows

SQL

- Teradata Database

Local Computing

Local Compute Contexts



Local Compute Contexts

Features

- Fastest performance for small data
- Limited parallelism
- IO-bound, not CPU-bound

Data Sources

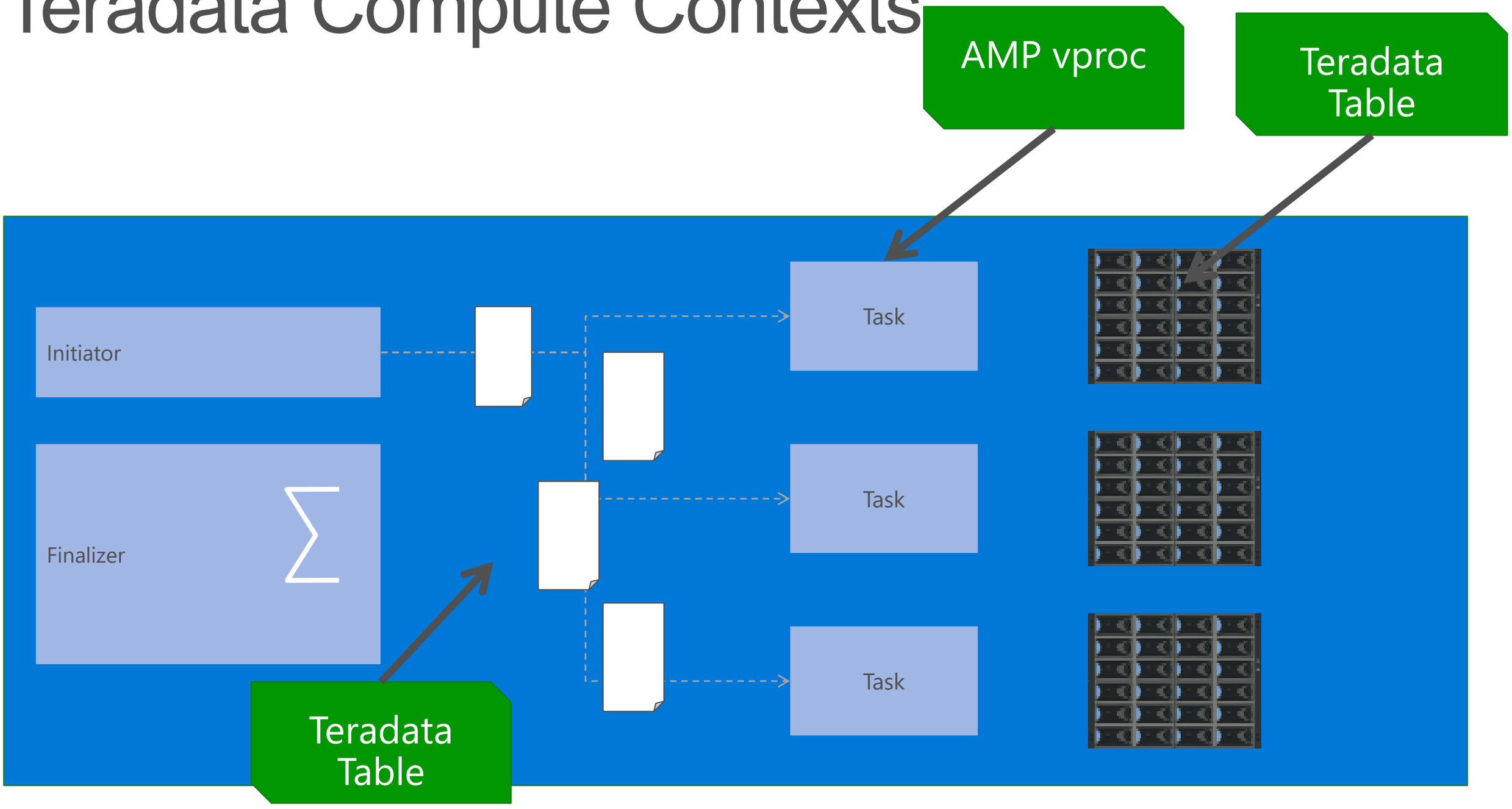
- ODBC connections
- Flat files (Spss, SAS, CSV, XDF)
- HDFS data sources

The XDF Format

- Compressed
- Chunkwise-columnar storage
- Optimized for R:
  - Data types
  - Memory-model and performance
- Enables 20+x speedups

ScaleR in Teradata

Teradata Compute Contexts



Teradata Compute Contexts

Features

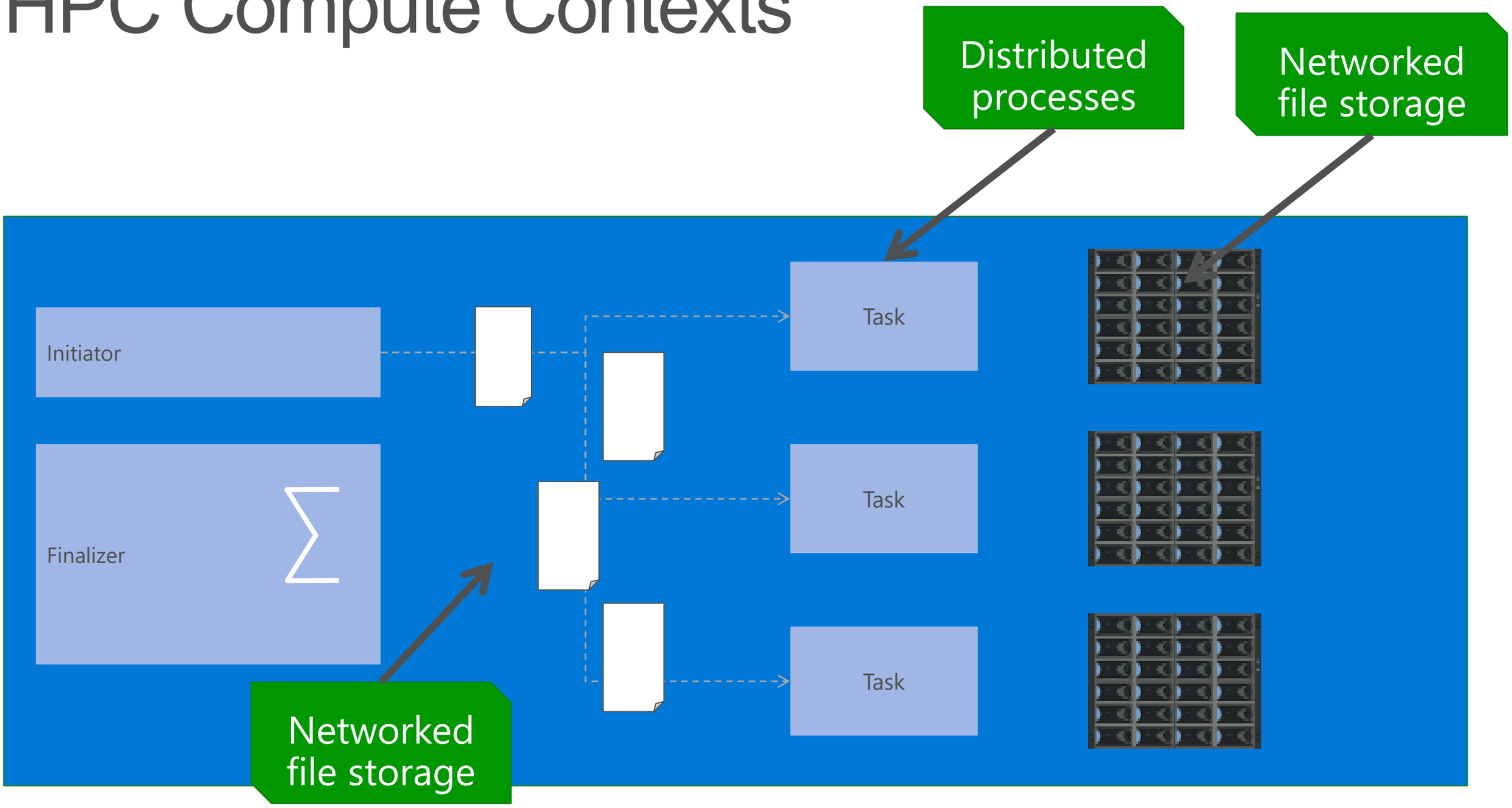
- Excellent performance
- Processes(vprocs) are shared resources
  - Resource contention
  - Custom transformations can have “memory leaks”

Data Sources

- Teradata data sources

ScaleR with HPC

HPC Compute Contexts



HPC Compute Contexts

Features

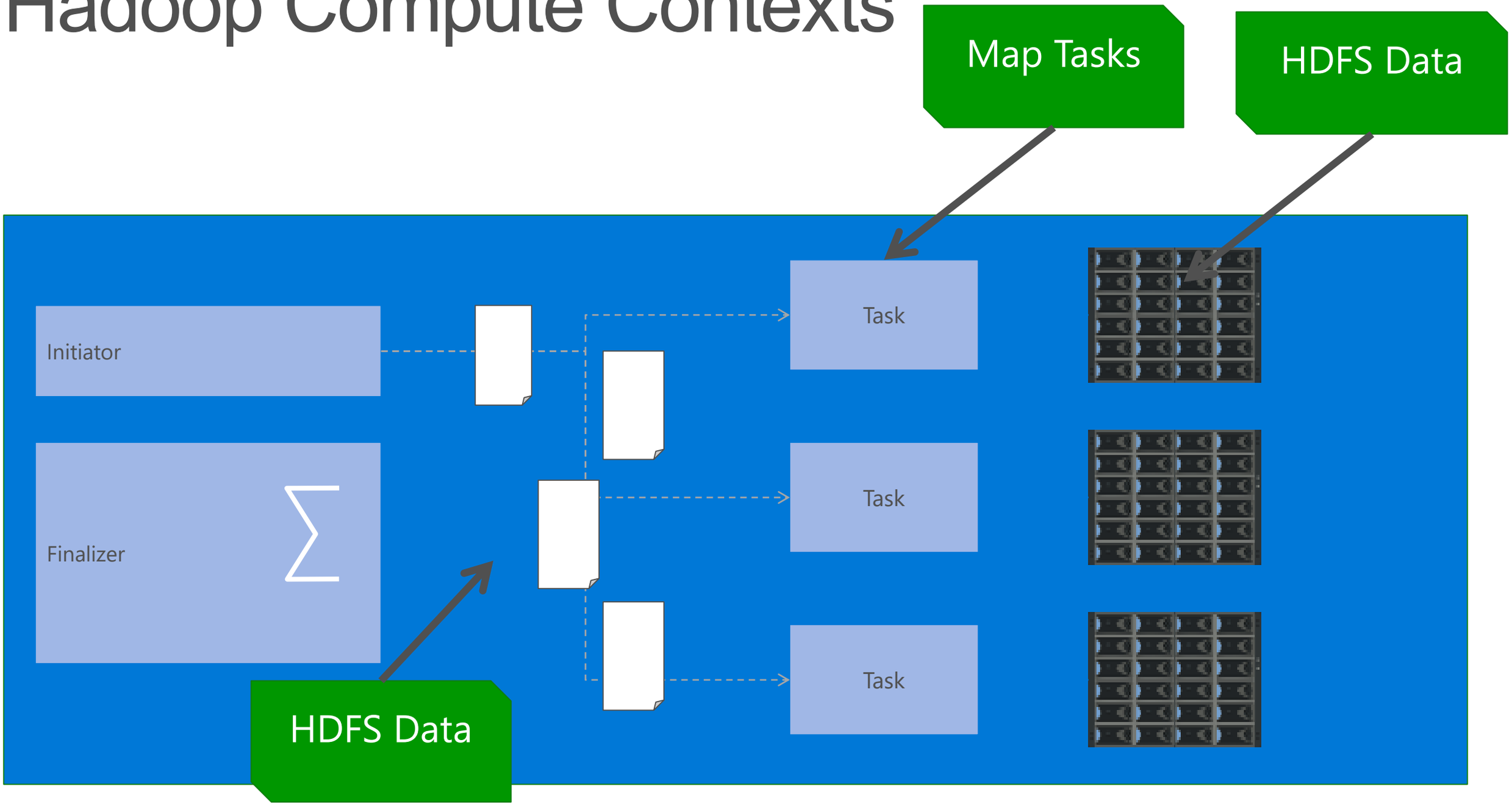
- Excellent performance
- Requires highly performance shared filesystems

Data Sources

- ODBC connections
- Flat files (Spss, SAS, CSV, XDF)
- HDFS data sources

ScaleR in Hadoop

Hadoop Compute Contexts



ScaleR Hadoop Configuration

- HDFS data source
- Local “share” directory
- HDFS “share” directory

Share directories must be writable by “you” from both the edge node and the data node.

RevoMPM: Multi-node Management

- Distributed shell
- Distributed copy
- Distributed and install R packages
- Deploy/Remove/Update Revolution R Enterprise
- Add/Remove users from nodes

Hadoop Compute Contexts

Features

- Slowest performance for multi-step tasks
- Unlimited parallelism
- Runtime dominated by MapReduce infrastructure

Data Sources

- HDFS data sources

Inside vs Beside with Hadoop

Inside Hadoop

- RRE installed on every data node
- ScaleR executes as MapReduce

Beside Hadoop

- RRE installed on an edge node
- ScaleR runs in parallel on a single system
- Connect to HDFS, local data, or databases

Coming Soon…

- ScaleR on Spark
- ScaleR on SQL Server

Questions?

Microsoft