REVOLUTION ANALYTICS

Real-Time Predictive Analytics with Big Data

from deployment to production

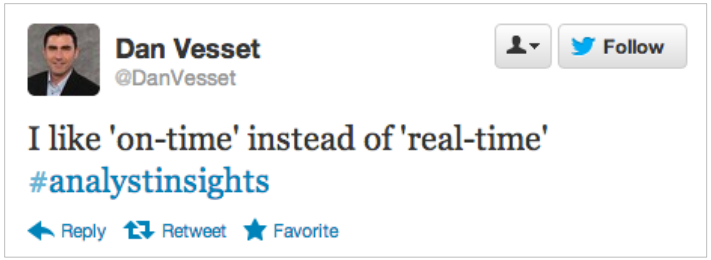
David M Smith @revodavid
VP Marketing and Community
Revolution Analytics

Today we'll discuss:

- What is "real-time predictive analytics with big data?"
  - Case study: UpStream Software
- Real-time big-data stack
- Five phases of deployment to production
- Just what do we mean by "Big Data" and "Real-Time", anyway?
- Recommendations
- Q&A

Real-Time Predictive Analytics with Big Data

- Real Time
  - Milliseconds? Seconds? Hours? Days?
  - In production, continuously updated
- Big Data
  - Size, flow rate, diversity
  - Conflict with ‘real time’
- Predictive Analytics
  - Rear view: description, aggregation, tabulation
  - Prediction, inference, statistical modeling



Case Study

www.upstreamsoftware.com

"Given that our data sets are already in the terabytes and are growing rapidly, we depend on Revolution R Enterprise's scalability and fast performance — we saw about a 4x performance improvement on 50 million records. It works brilliantly."

John Wallace, CEO Upstream Software

From: "How Big Data is Changing Retail Marketing Analytics"
http://bit.ly/upstream-webinar

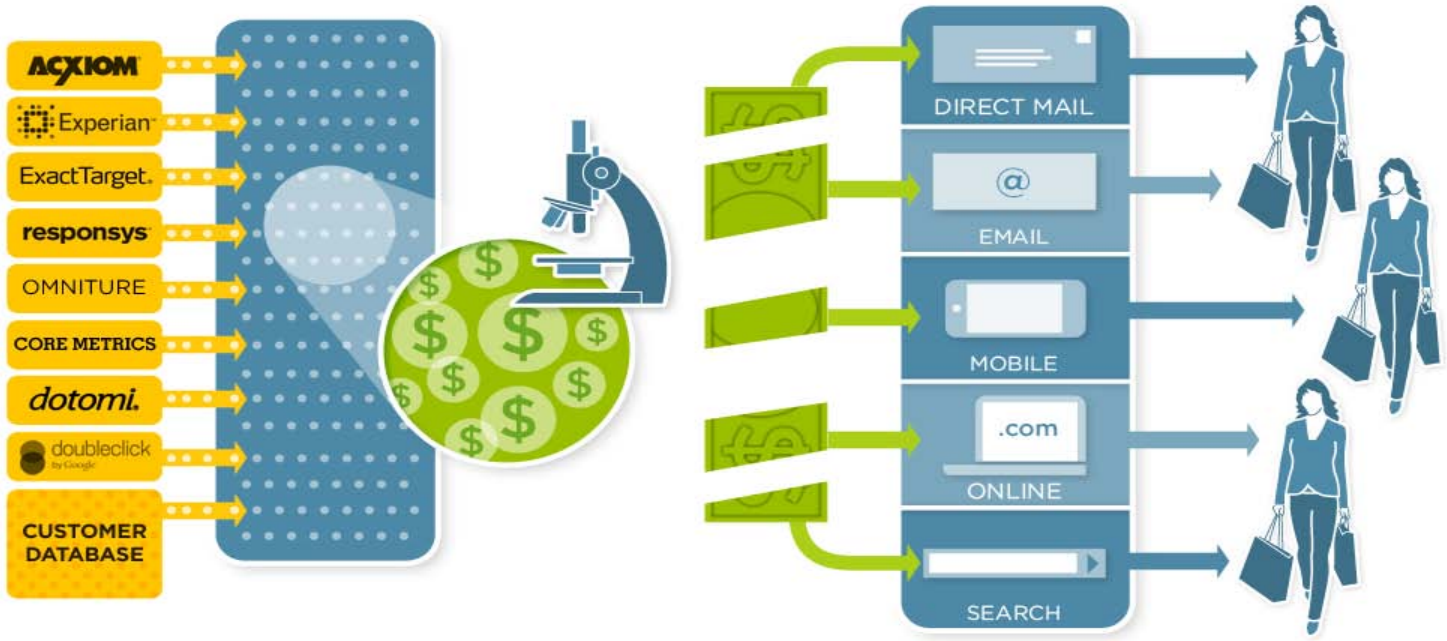
REVOLUTION ANALYTICS

UpStream Software

UpStream's Big Data Analytics engine analyzes and optimizes marketing mix for UpStream's retailer clients

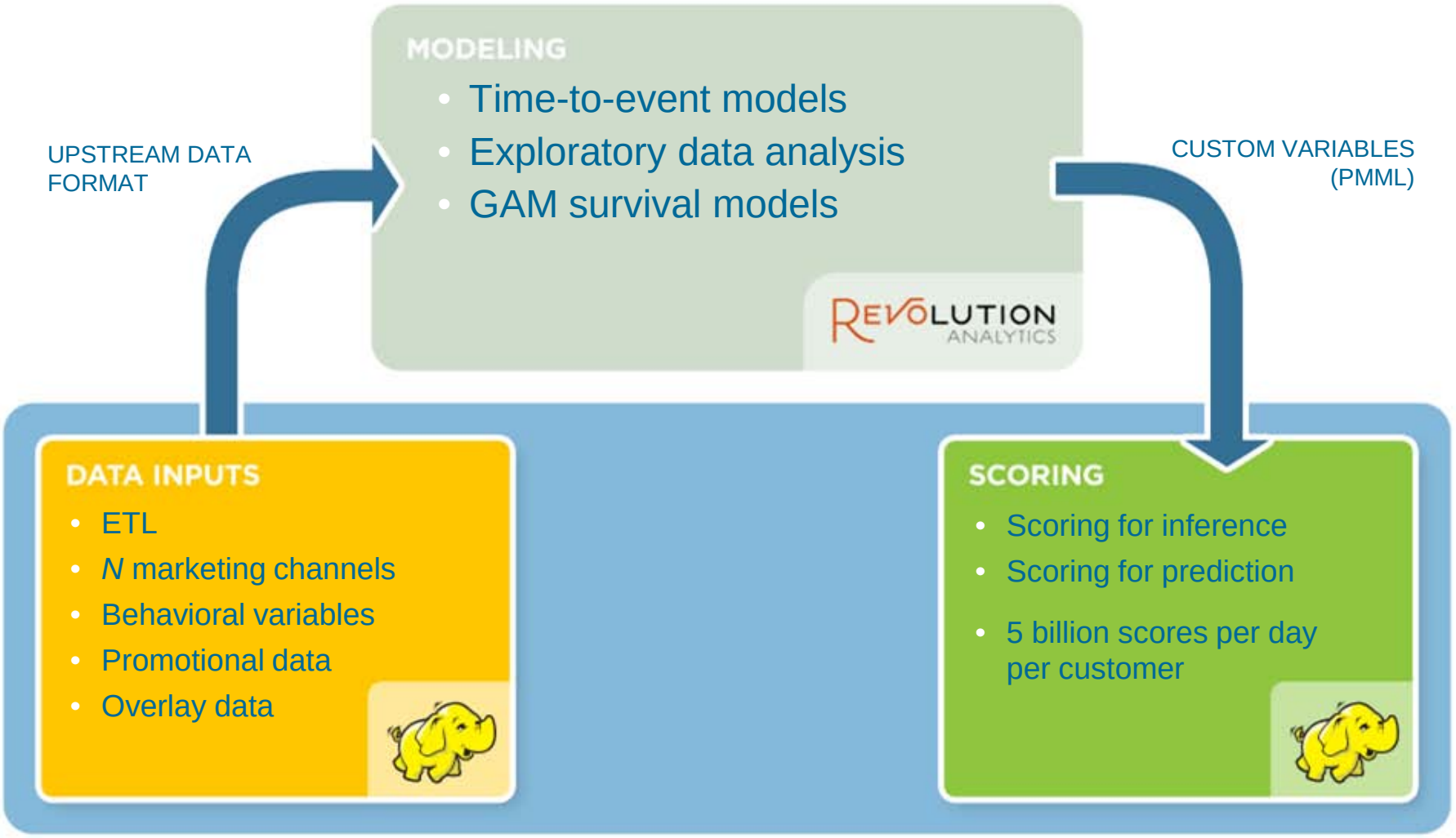
- Customer analytics and segmentation
- Revenue/ action attribution among channels and marketing programs
- Customized Next Best Action per individual prospect or customer
- Event-Triggered Marketing: responses to consumer actions

Big Data



- Demographics: consumer, product, market
- Actions: web clicks, email clicks, mobile app usage, call center logs, social, search …
- Outcomes: impressions, touches, orders (retail, online, mobile)

Predictive Analytics

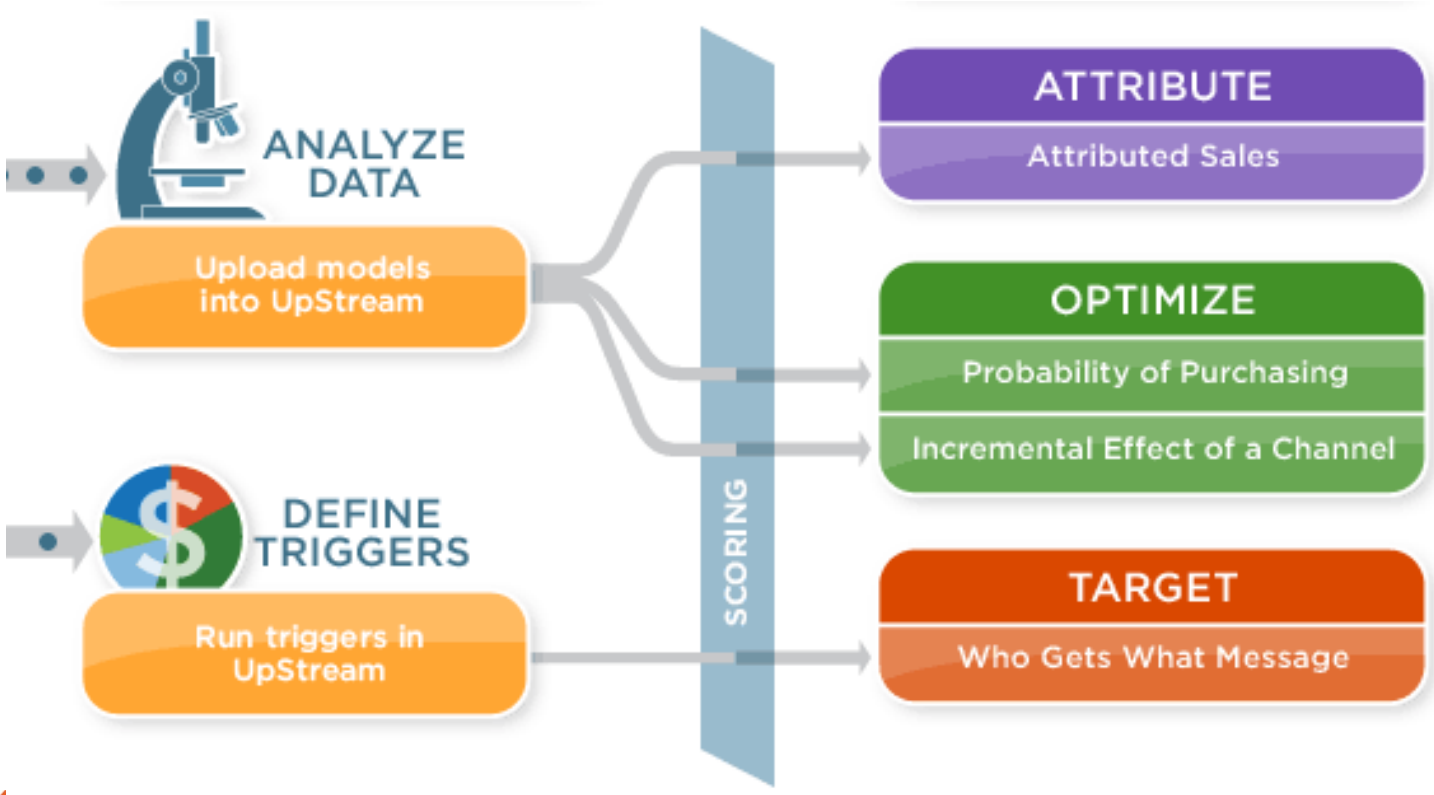


Real Time

InfoWorld Big Data

Williams-Sonoma uses big data to zero in on customers

To target individual customers, Williams-Sonoma needed data from a broad swath of sources, a Hadoop platform, and a dashboard to make sense of it all

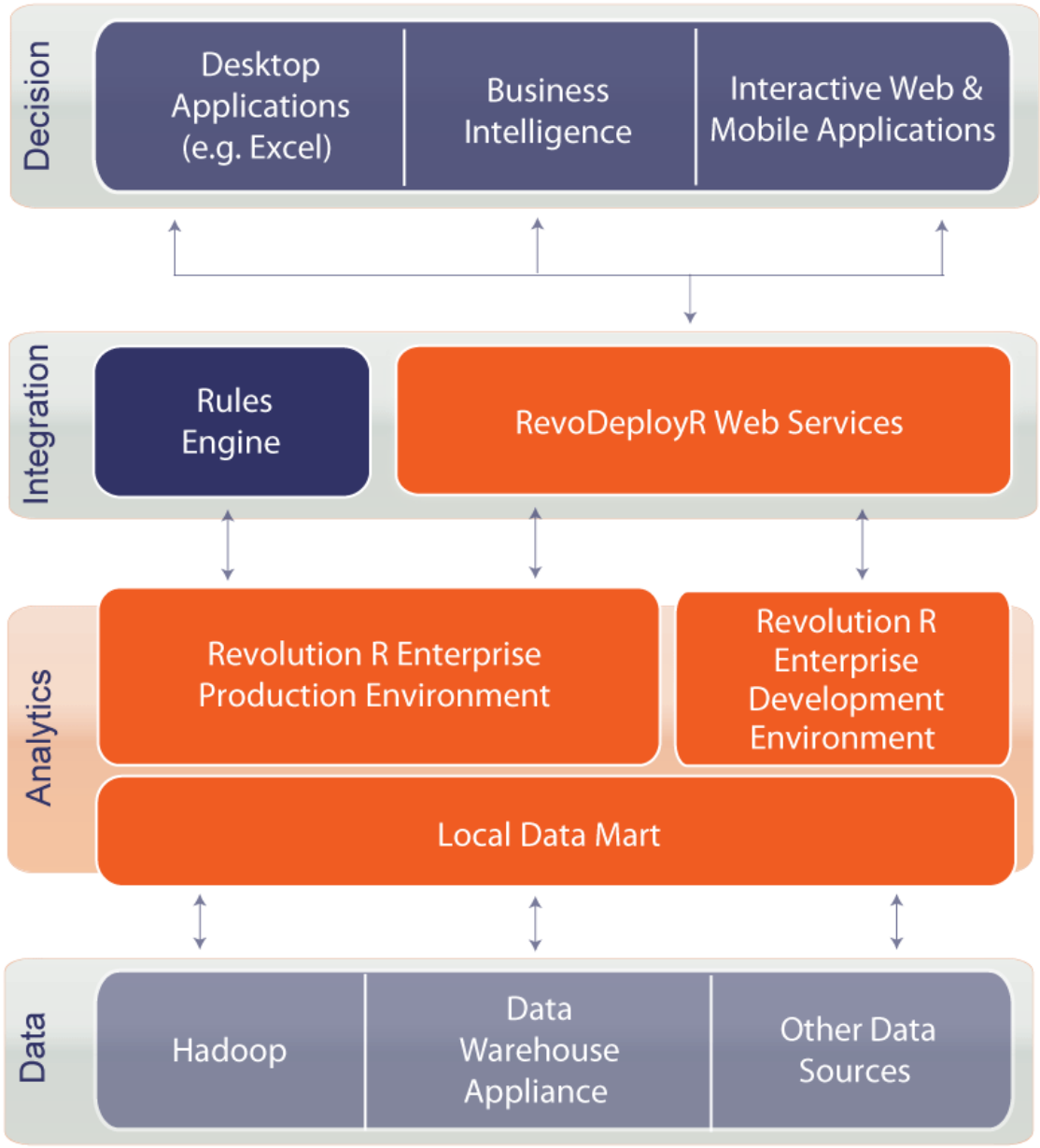


http://www.infoworld.com/d/big-data/williams-sonoma-uses-big-data-zero-in-customers-206641
November 8, 2012

Predictive vs Descriptive Analytics

- Retail sales are influenced by stores near the home
- Newer customers are more sensitive to marketing
- Google keywords perform worse than you think
- Display advertising performs better than you think
- Online sales benefit more from seasonal events than retail stores

Real-time Big Data Predictive Analytics Stack



REVOLUTION ANALYTICS

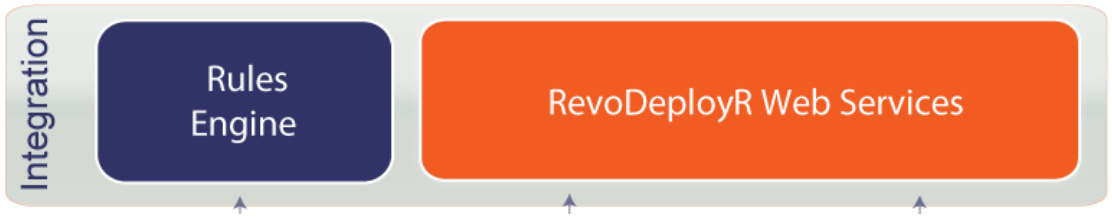
Data Layer

- Structured data
  - RDBMS, NoSQL, Hbase, Impala
- Unstructured data
  - Hadoop MapReduce
- Streaming data
  - Web, social, sensors, operational systems
- Some *descriptive* analytics done here

Analytics layer

- **Predictive** analytics technology
  - Development environment: build models
  - Production environment:
    - **Deploy** real-time scoring
    - **Engine** for dynamic analytics
- Local data mart
  - Static, periodically updated from data layer
  - Improves performance

Integration Layer



- Connective tissue between end-user applications and analytics engine
- Engines for flow control, **real-time scoring**
  - Rules engine / CEP engine
- API for Dynamic Analytics
  - Brokers communication between app developers and data scientists

Decision Layer

- End-user interface to analytics system
  - Customers
  - Operations
  - Business analysts
  - C-suite
- Familiar & simple interfaces
  - Supports a range of end-user technologies
  - Level of UI complexity dictated by need

Real-Time Deployment

Five phases of deploying real-time predictive analytics with big data to production:

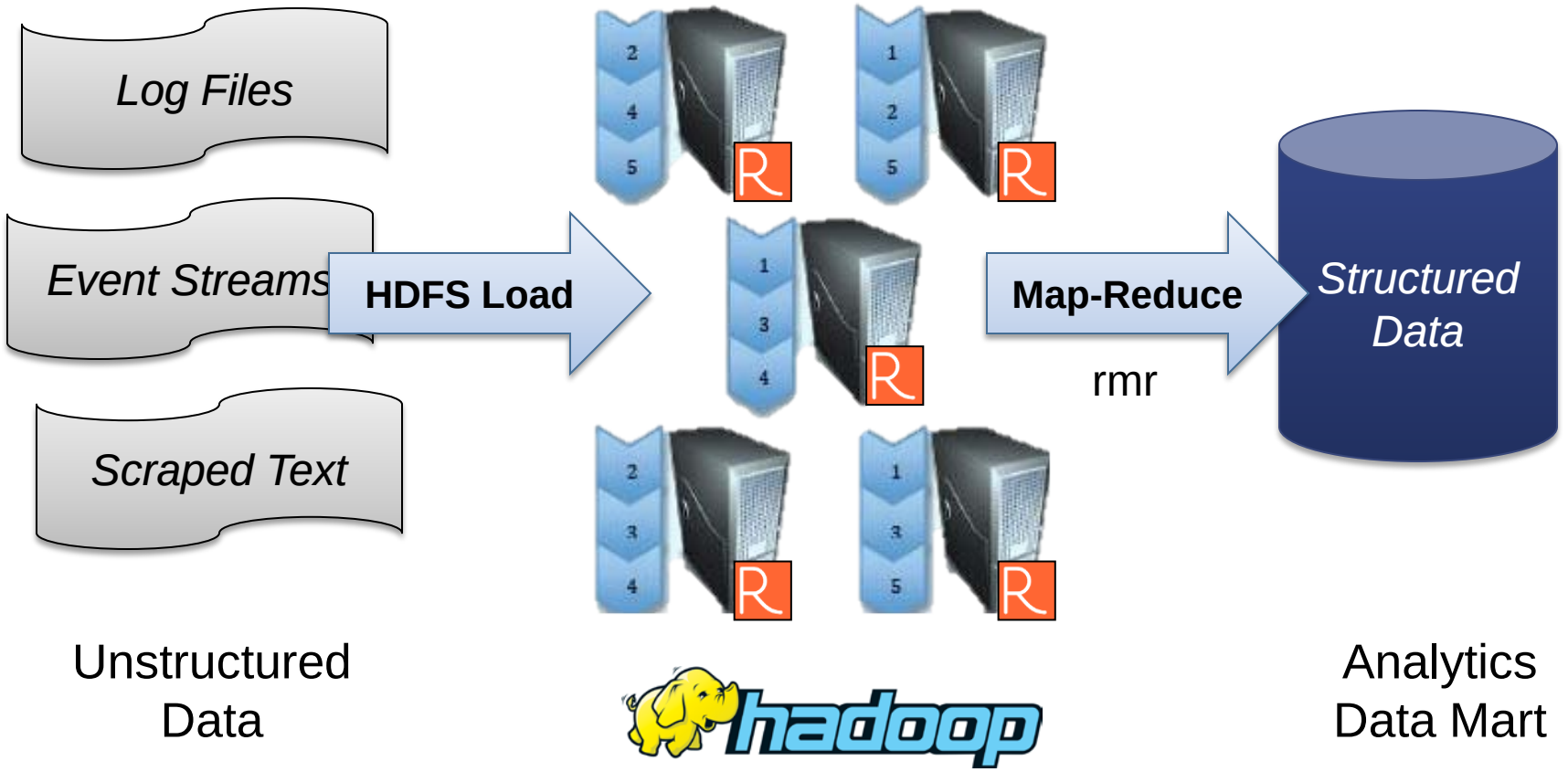
1. Data Distillation
2. Model Development
3. Validation and Deployment
4. Real-time Scoring
5. Model refresh

Phase 1: Data Distillation

The data in the Data Layer isn't yet ready for predictive modeling. We need to:

- **Extract** features from unstructured text
  - Topic modeling, sentiment analysis
- **Combine** disparate data sources
- **Filter** for populations of interest
- **Select** relevant features and outcomes for modeling
- **Export** to the data mart for predictive modeling

Example: Data Distillation in Hadoop

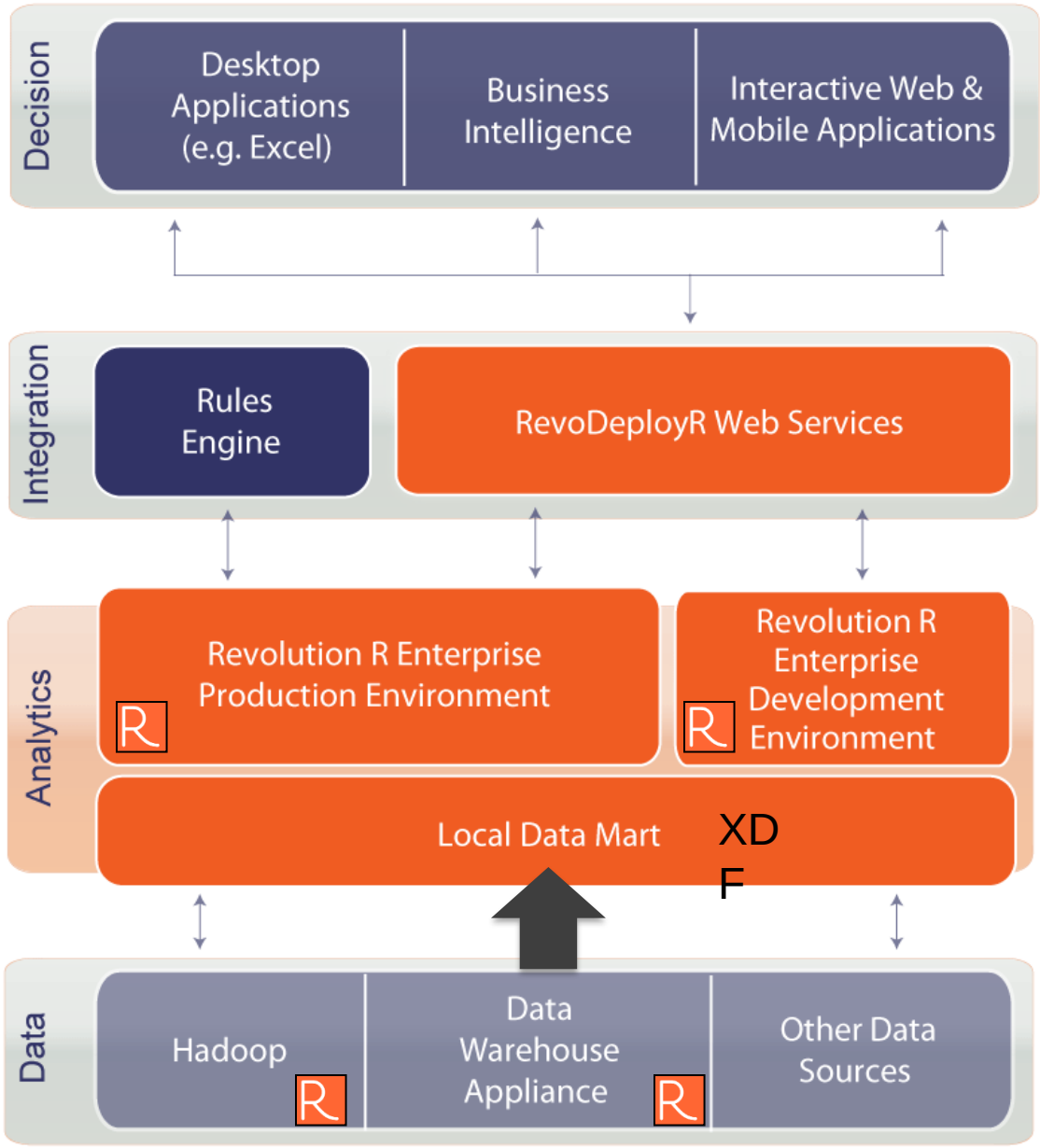


Revolution R Enterprise for Hadoop: bit.ly/r-hadoop

Automating the Extraction Process

- This process needs to be **repeatable** and **maintainable**
- Create a re-usable R script:
  - **ASCII**: rxDataStep
  - **SQL**: rxImport, ROracle, RMySQL
  - **NoSQL**: Rcassandra, rmongodb
  - **Hadoop**: rmr, rhbase, Rhive
  - **Appliances**: nza, teradataR, PL/R
  - **Feeds**: rjson, XML, RCurl, twitteR, python
- R script runs from Analytics Layer
  - Processing: in Data Layer
  - Output: structured file for Data Mart

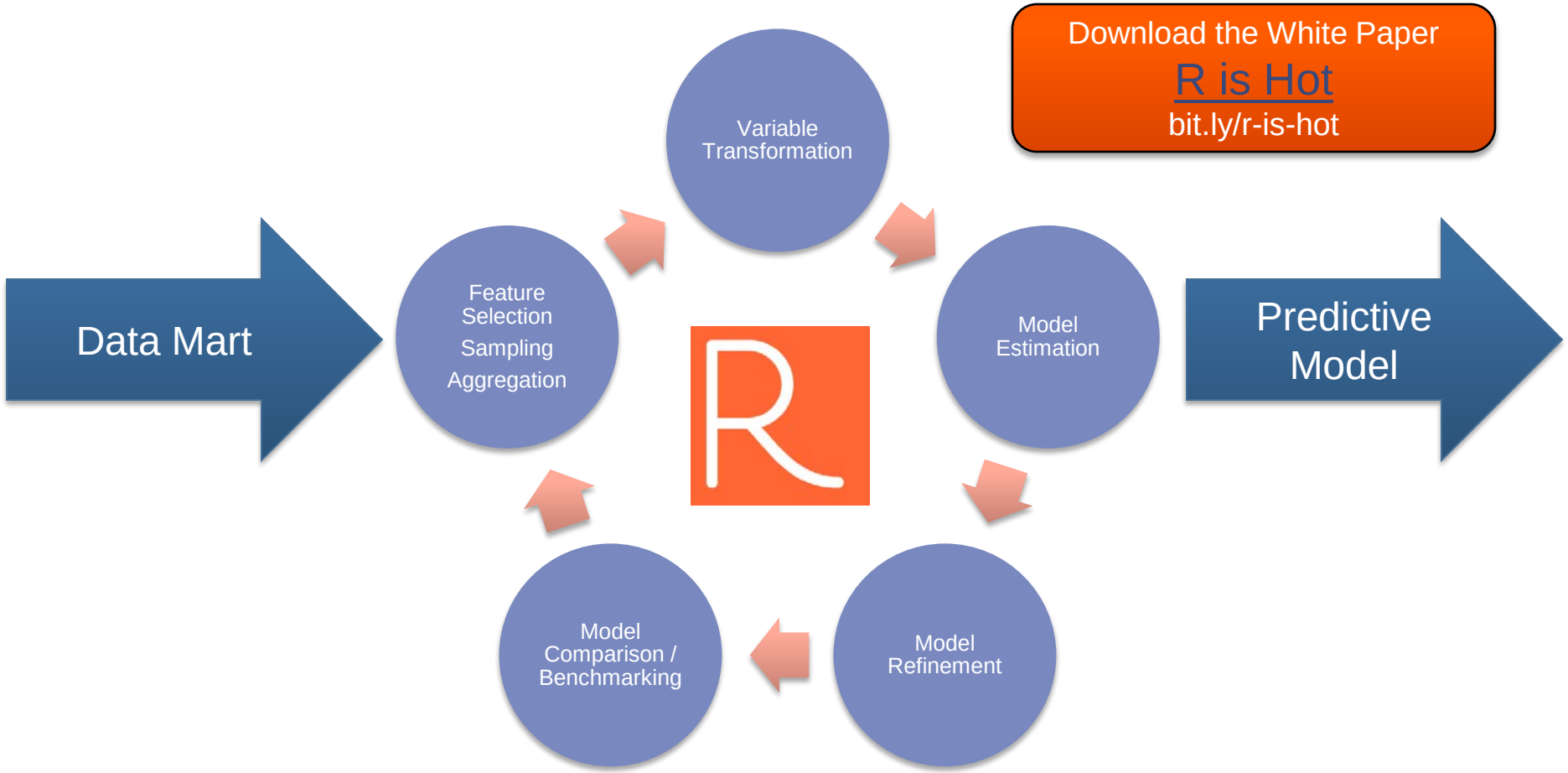
Data Distillation Process



Phase 2: Model development

- **Goal**: create a predictive model that is
  - Powerful
  - Robust
  - Comprehensible
  - Implementable
- Key requirements for Data Scientists:
  - Flexibility
  - Productivity
  - Speed
  - Reproducibility

The Model Development Cycle



Technology considerations

- Moving Big Data is slow
  - So just move it once, to data mart near the development and production environments
- Static data needed for modeling cycle
  - But have a plan to refresh and update
- Data layer not optimized for *predictive* analytics
  - But good for descriptive (data distillation)
- Revolution R Enterprise optimizations for the analytics layer
  - R; XDF file format; Parallel External Memory Algorithms

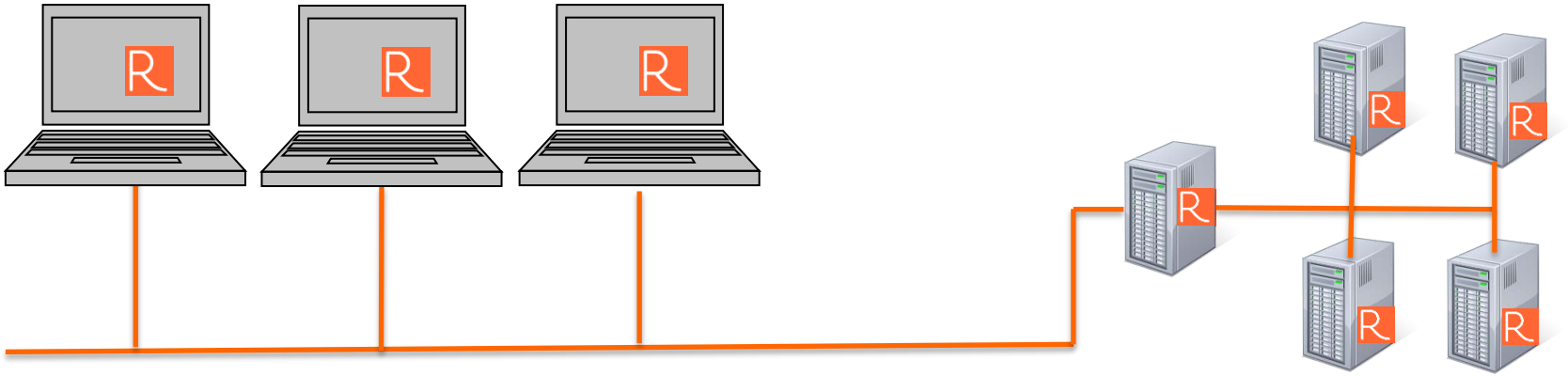
Predictive Modeling Algorithms

Algorithm	Example Applications	Big Data
Data Step	ETL, data distillation, record/variable selection, variable transformation	✓✓
Descriptive Statistics	Exploratory Data Analysis, Data Validation	✓✓
Tables & Cubes	Reporting, contingency analysis	✓✓
Correlation / Covariance	Factor Analysis, Value at Risk	✓✓
Linear regression	Forecasting, Net present value estimation	✓✓
Logistic Regression	Response modeling, offer selection	✓✓
Generalized Linear Models	Capital reserve estimation, climate modeling	✓✓
K-means clustering	Customer Segmentation	✓✓
Decision Trees	Dynamic pricing, classification, variable importance	✓✓
Model Prediction	Real-time Scoring (decisions, offers, actions)	✓✓
R CRAN packages	Everything else	
Parallel & distributed computing with R	Simulations, By-Group analysis, ensemble models, custom applications	✓

High performance with distributed clusters

Data Scientists / Modelers

Grid computing cluster



Revolution R Enterprise

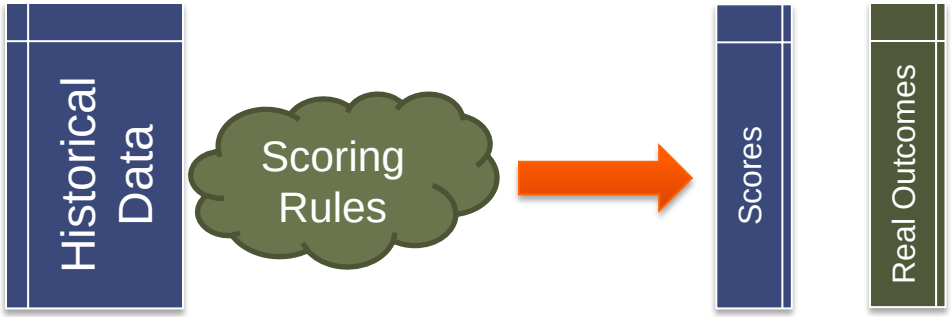
Platform LSF
Microsoft HPC Server
Microsoft Azure
Ad-hoc grids (SNOW)
Shared SMP server
Hadoop / HDFS

Phase 3: Model Validation and Deployment

- **Scoring rules** map parameters to outputs
  - Parameters: information **known** at real time
    - prospect ID, product, webpage, …
  - Scores: values **inferred** in real time
    - prices, recommendations, actions, …

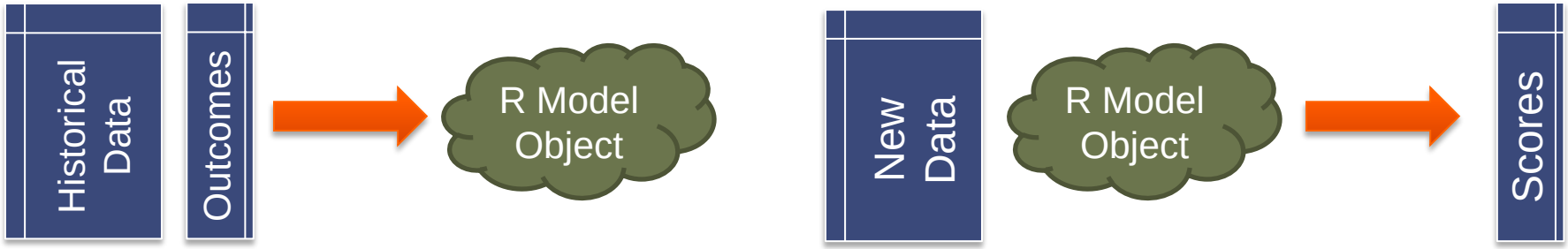
- Validation: backtest with historical data
- Deployment: code running in real time

Validation for production



- Refresh data mart
- Rebuild model, withholding a validation set
  - Random sampling
- Create accuracy measure
  - ROC, positive rate, false positive rate
- Measure and monitor

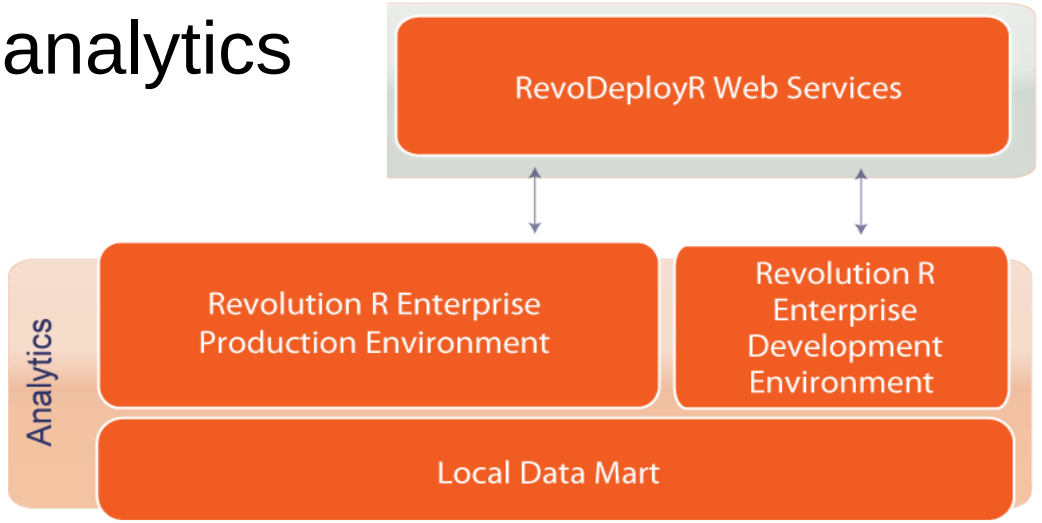
Deployment with R code



- Scoring rules captured as “R model objects”
- Move R code directly from development environment to production server
  - No recoding: Lowest cost, greatest speed & reliability
  - Most flexibility (not limited in model choice)
  - Easy to validate with custom accuracy measures

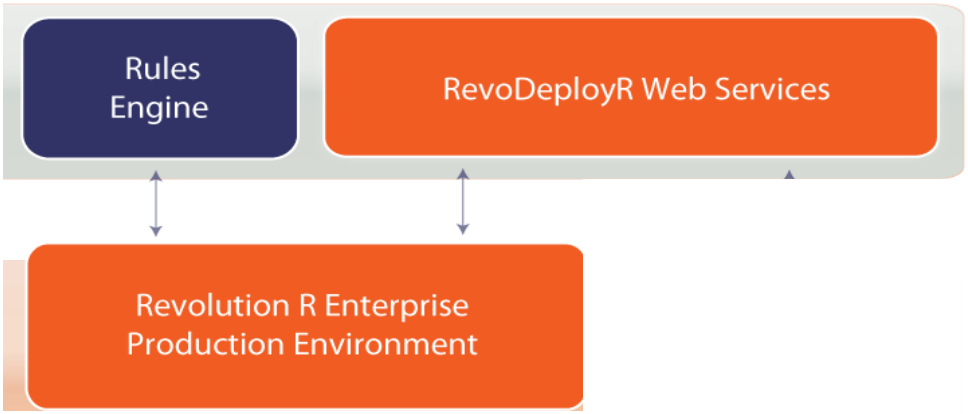
Managing and Scaling Deployed R Code

- Revolution R Enterprise includes RevoDeployR
  - Code management & tracking
  - Security & resource allocation
  - Scalability for real-time demands
  - Web Services API
- Scoring *and* dynamic analytics
  - Data visualization
  - custom reporting
  - Ad-hoc analytics
  - Interactive data apps



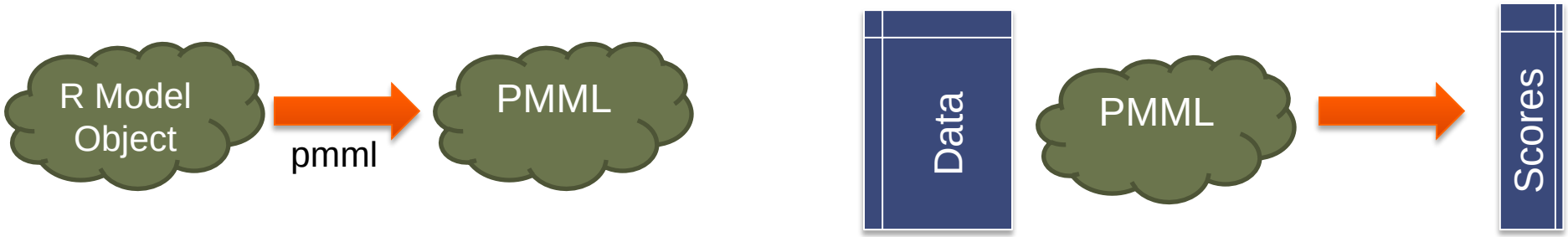
Code Conversion for Scoring

- Business rules (e.g. IBM ILOG)
  - Create directly with R code from model object
- Recode in other languages
  - SQL
  - C++, etc.
- Low latency
- Slow and costly deployment
- Potential for errors



Scoring via PMML

- Some predictive models can be expressed in the PMML standard (www.dmg.org)
  - Not all, but growing all the time

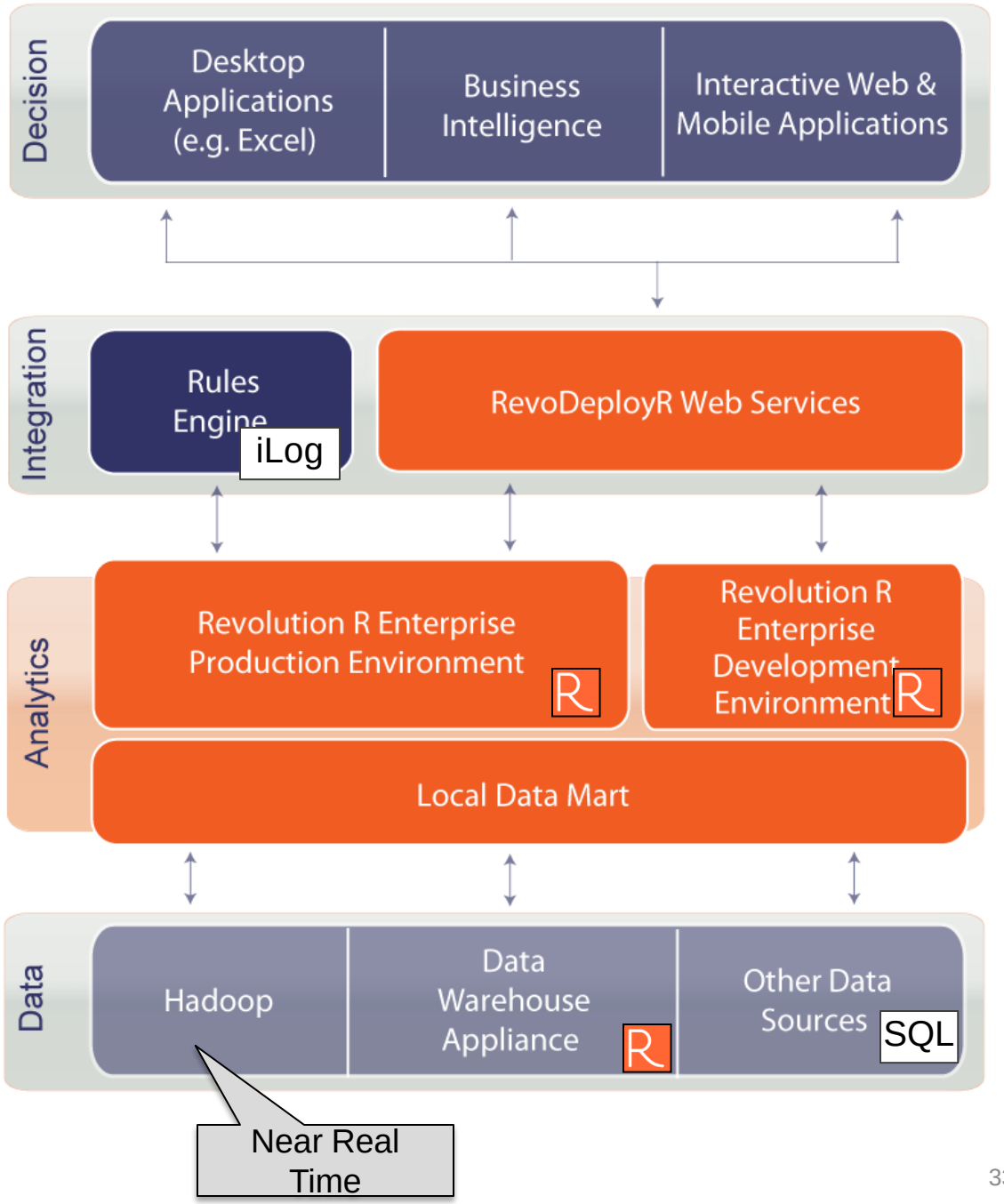


- Easily generate PMML with R “pmml” package
- PMML Scoring Engine: Zementis Adapa
- Databases and appliances may support PMML
  - But supported standards vary

Phase 4: Real-Time Scoring

- Scoring triggered by Decision Layer, brokered by Integration Layer
  - R code
    - Revolution R Enterprise Server
    - In-appliance (e.g IBM PureData System for Analytics)
  - SQL
    - In-database
  - Rules
  - Compiled code
    - Bespoke engines
- May be using **hardware** from data layer
  - But not the actual data

Real-Time Scoring: review



Phase 5: Model Refresh

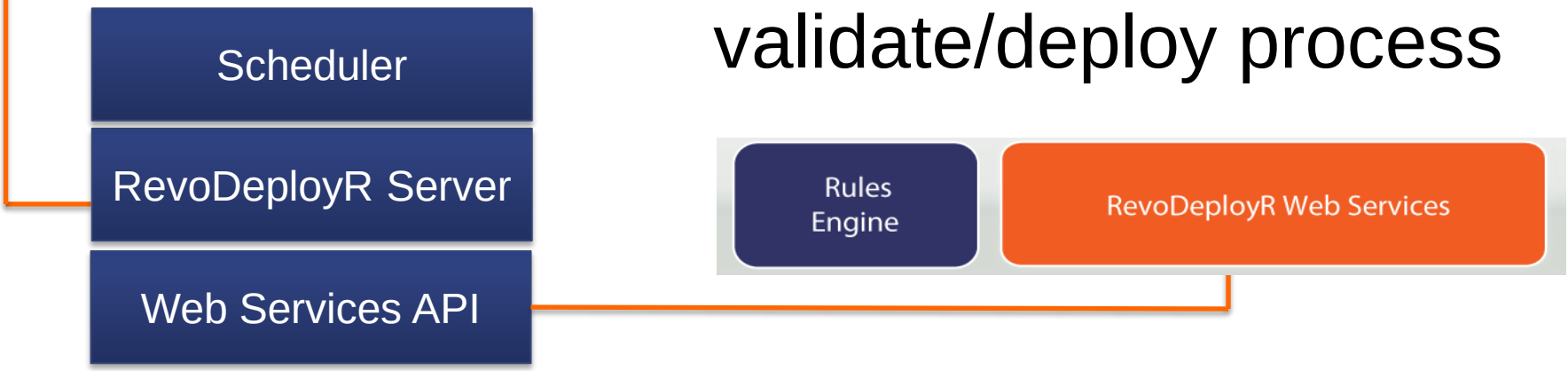
- Deployed scoring rules no longer connected to Data Layer or Data Mart
  - Enables real-time performance
  - Need to refresh using recent data

- Model refresh process
  - Use data extract script
  - Re-run model script
  - Statistical review OR automated validation

Scheduled Batch Updates

Revolution R Enterprise
Production Server Cluster

- Periodic model refresh
  - Weekly
  - Daily
  - Hourly
- Automated validate/deploy process



Re-developing the model

- Even then, the underlying structure of the data may change:
  - Important variables become non-significant
  - Non-significant variables become important!
  - New data sources become available
- Model accuracy drift is a warning sign
- Return to Phase 2 or Phase 1

How big is 'Big data'?

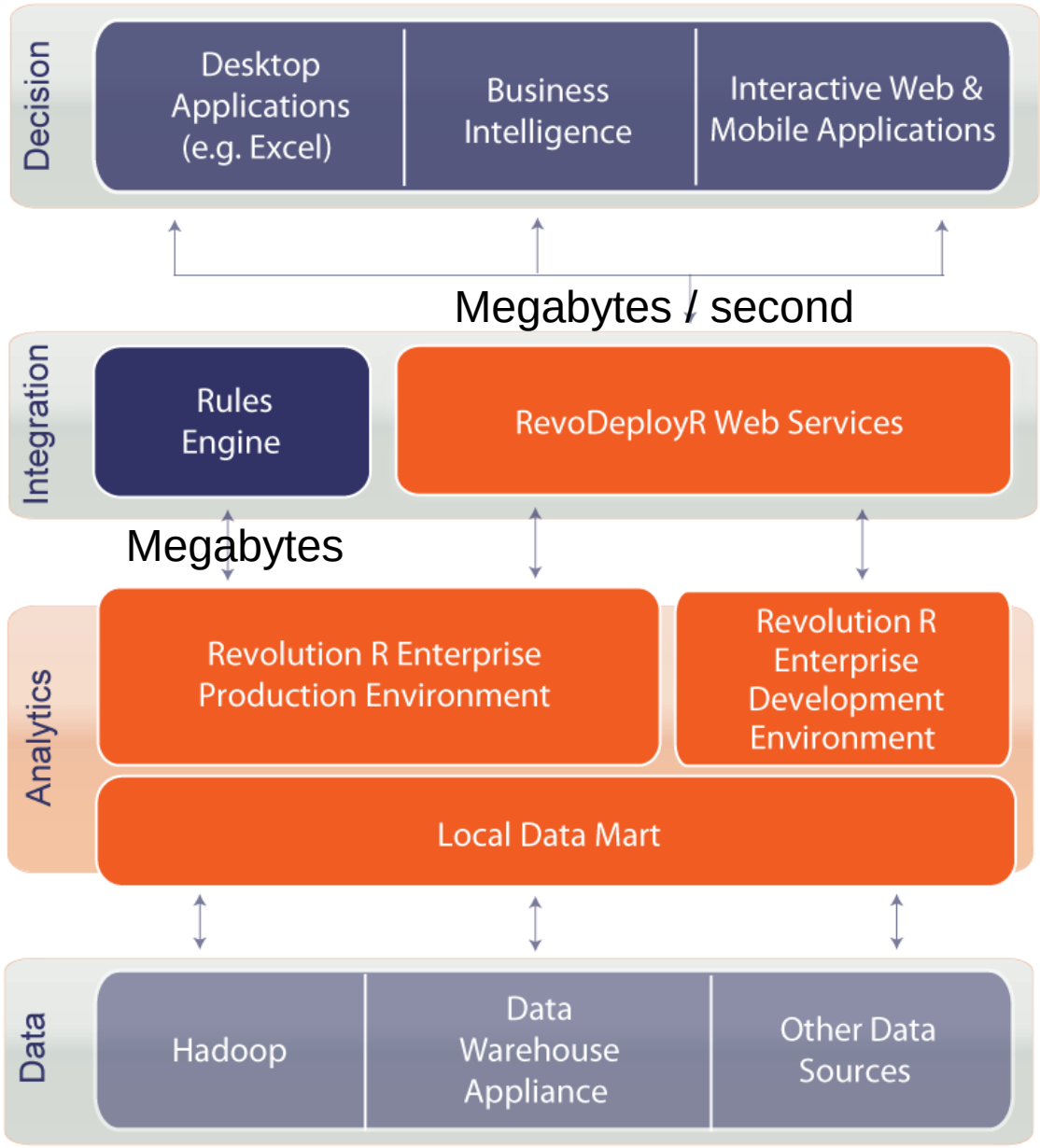
Kilobytes / second

Megabytes / second

Megabytes

Gigabytes ➔ Terabytes

Petabytes ➔ Exabytes



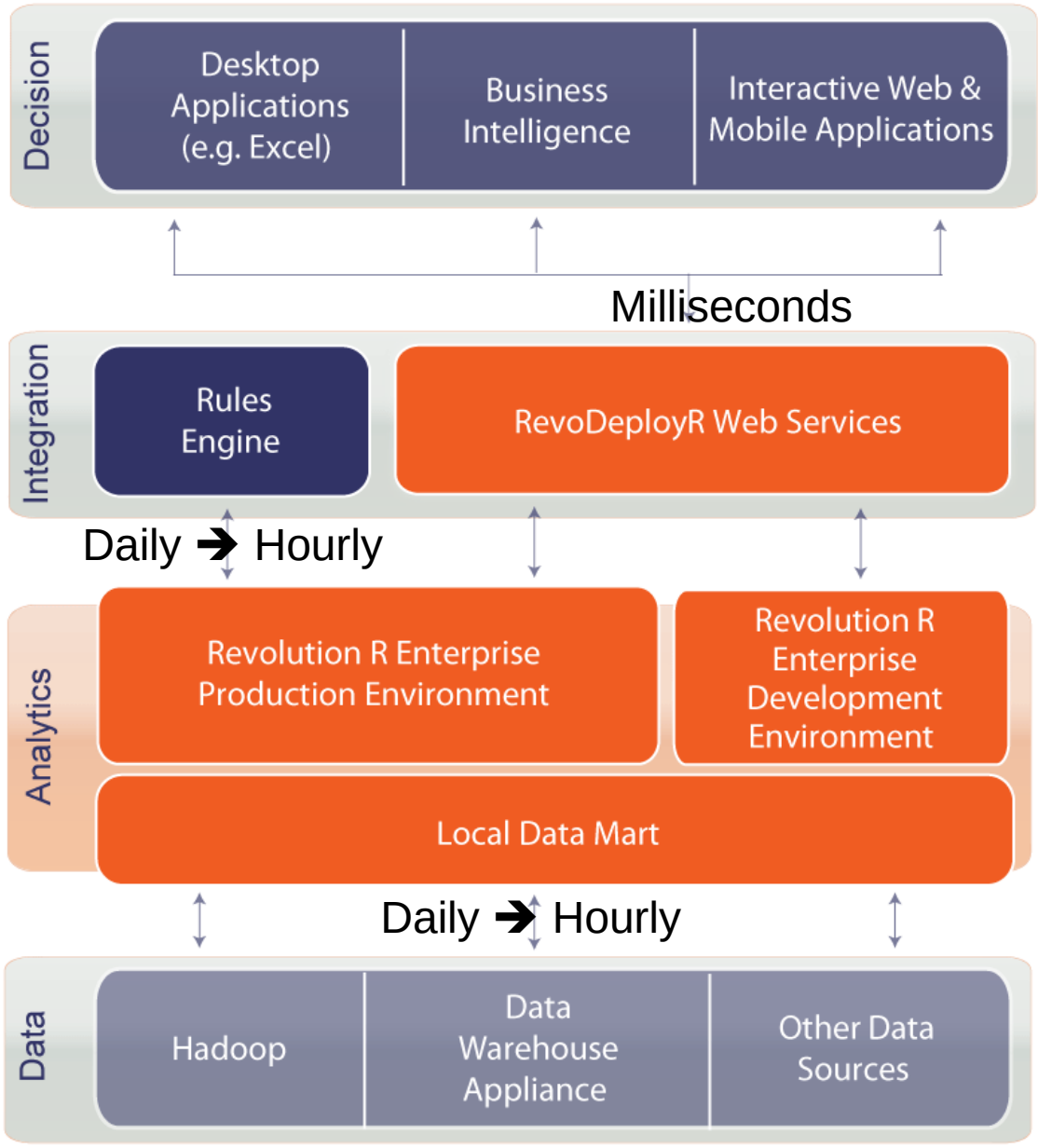
REVOLUTION ANALYTICS

How fast is 'Real time'?

Seconds or less

Hours ➔ Minutes

Weeks ➔ Days



Recommendations

- Architect your Predictive Analytics Stack
  - Best of breed vs single-vendor stack
  - Diversify data sources and decision apps
- R = Predictive Analytics

  Revolution R Enterprise provides:
  - Analytics Layer: big data, performance
  - Integration Layer: scalability, reliability
- Get Help to Get Started
  - Consider an SI partner for architecture, best practices and implementation advice

Resources

- Revolution R Enterprise : R for Big Data
  - www.revolutionanalytics.com/products
- Revolution Analytics Consulting Services
  - www.revolutionanalytics.com/services
  - Contact us: bit.ly/hey-revo
- Rhadoop : Connecting R and Hadoop
  - bit.ly/r-hadoop

- Contact David Smith
  - david@revolutionanalytics.com
  -  @revodavid
  - blog.revolutionanalytics.com

Thank you.

The leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com | 650.646.9545 | Twitter: @RevolutionR