Difinity

Microsoft Data Platform Business Intelligence Analytics Conference

Auckland, New Zealand
12-14 February 2018

www.difinity.co.nz

Microsoft Difinity

Scalable Data Science with SparkR on HDInsight

Lace Lofranco
Senior Software Development Engineer
Microsoft

Survey

Agenda

- Spark Fundamentals
- Spark on HDInsight
- Overview of R and SparkR
- Machine Learning in SparkR
- UDFs in SparkR

Spark Fundamentals

hadoop MapReduce

HIVE

STORM

TEZ

mahout

Apache Flink

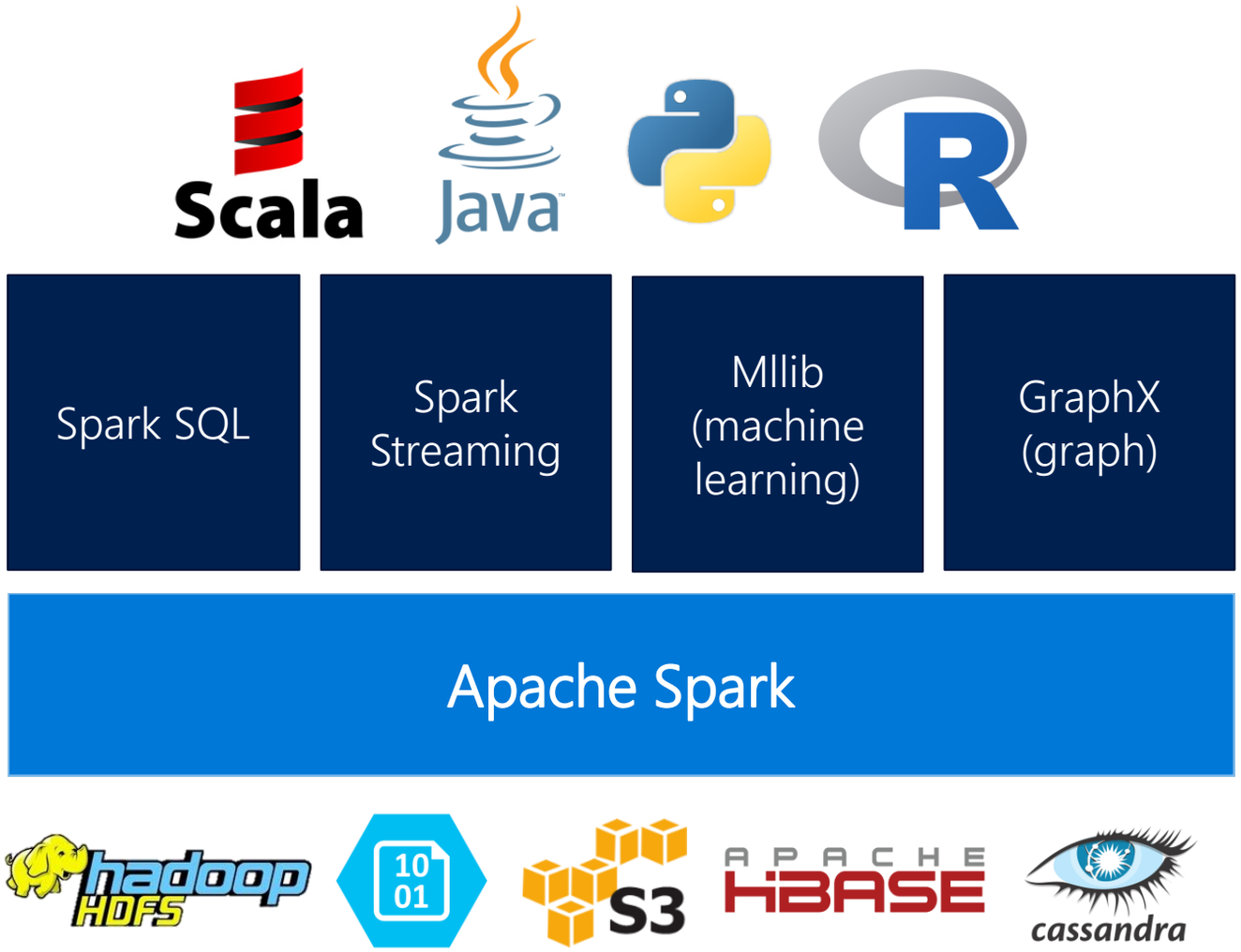
FLUME

APACHE GIRAPH

oozie

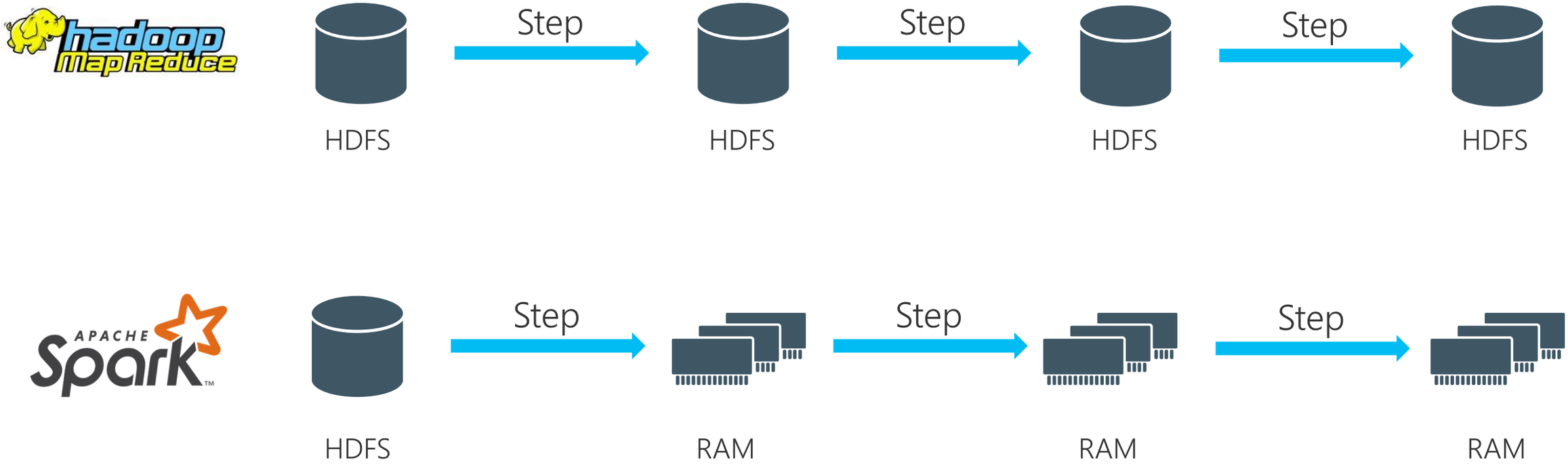
Apache Spark

a **unified computing engine** and a set of libraries for parallel data processing on computer clusters

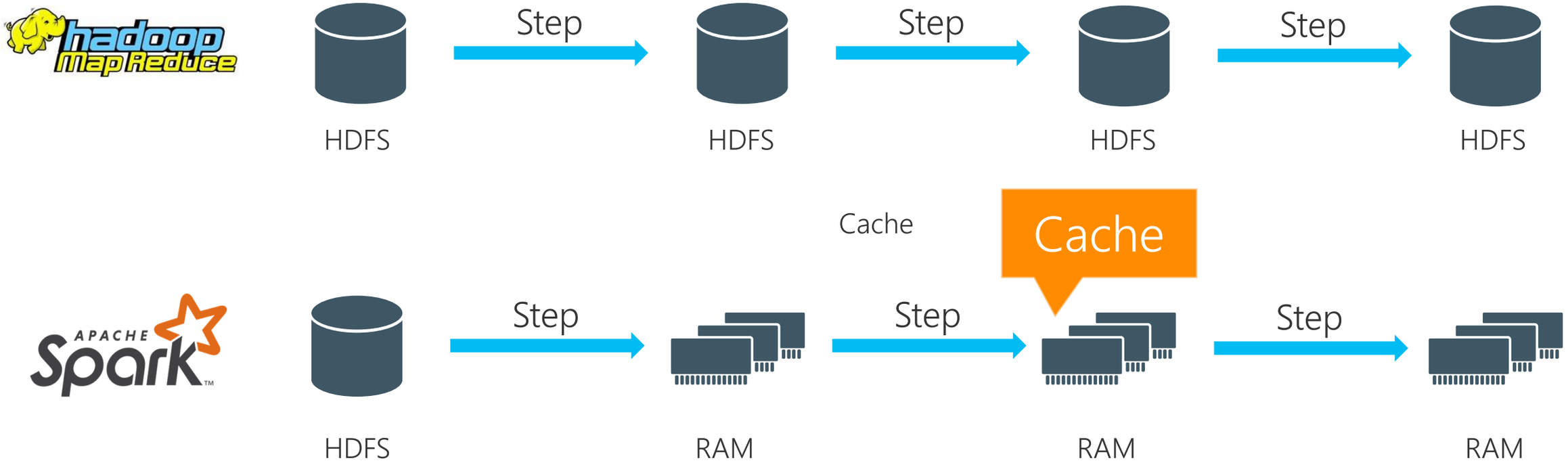


Spark: The Definitive Guide, Matei Zaharia, Bill Chambers

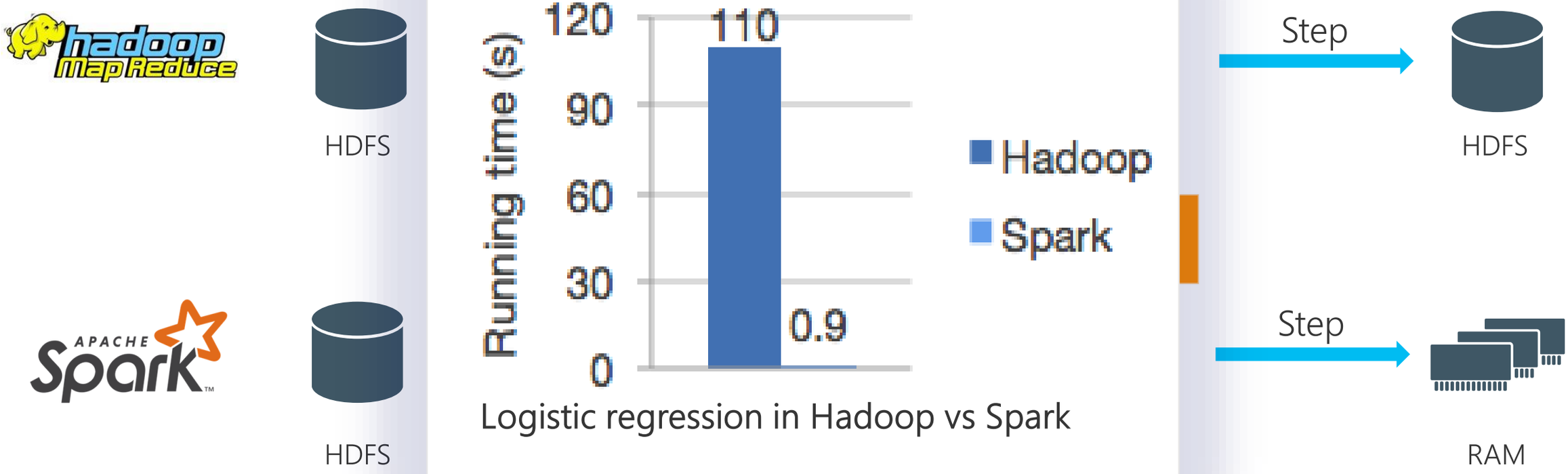
Why Spark is fast



Why Spark is fast



Why Spark is fast



Source: http://spark.apache.org/

Apache Spark: APIs

RDDs

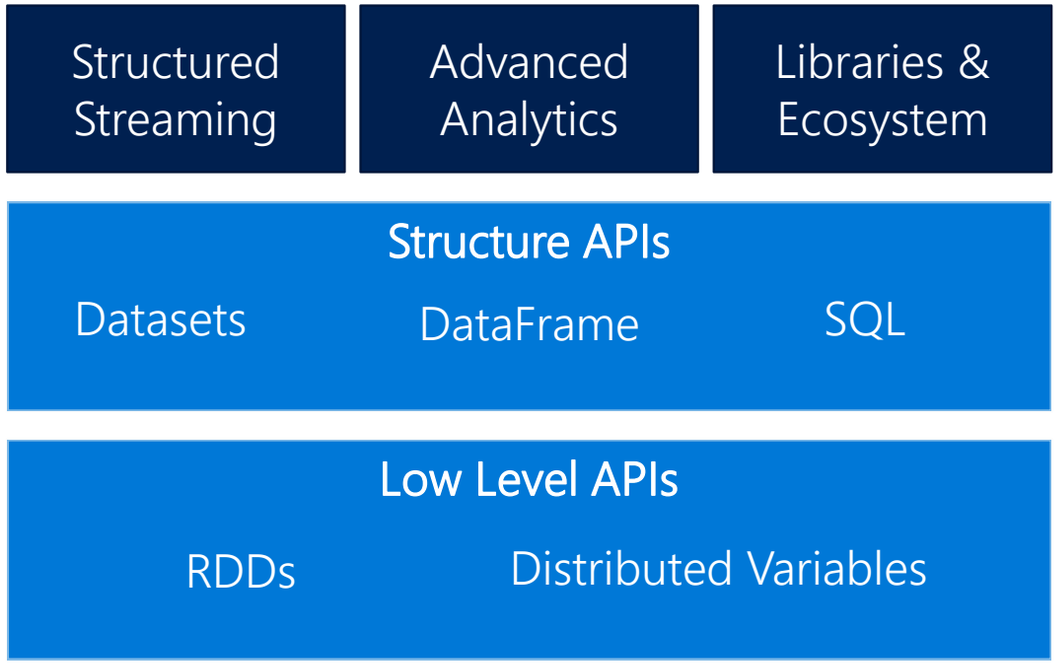
Core building block of data processing pipelines

DataFrames

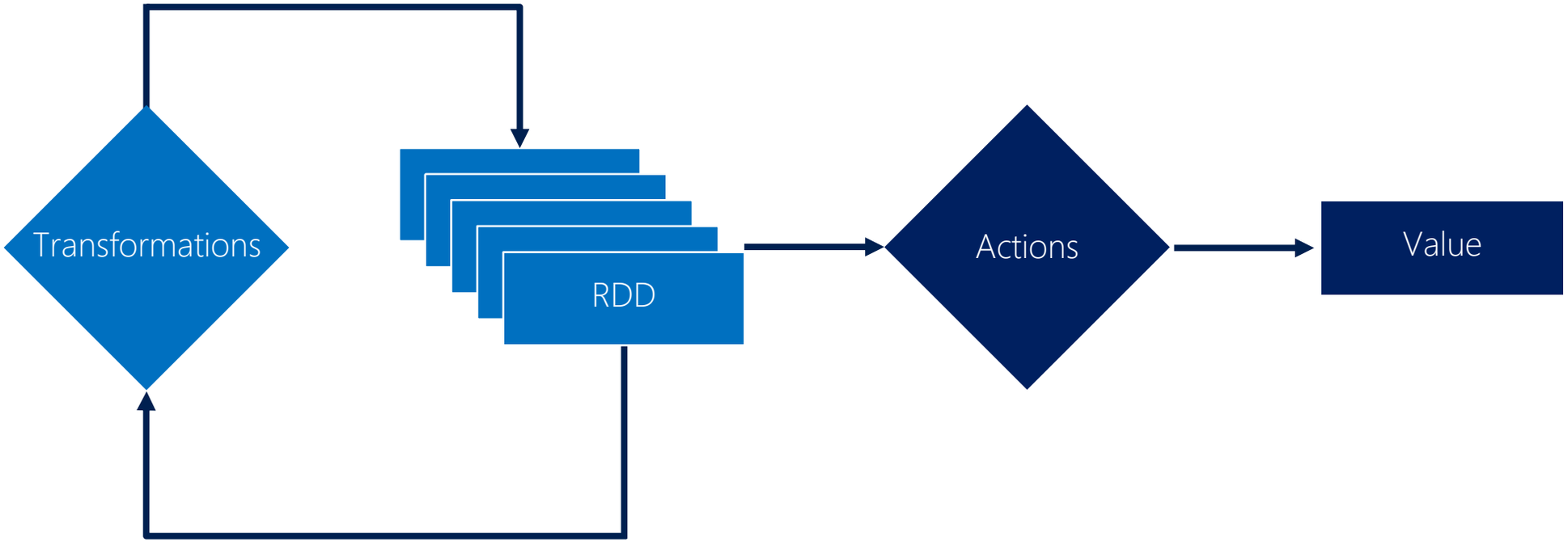
High level APIs that take advantage of query optimizer

Datasets

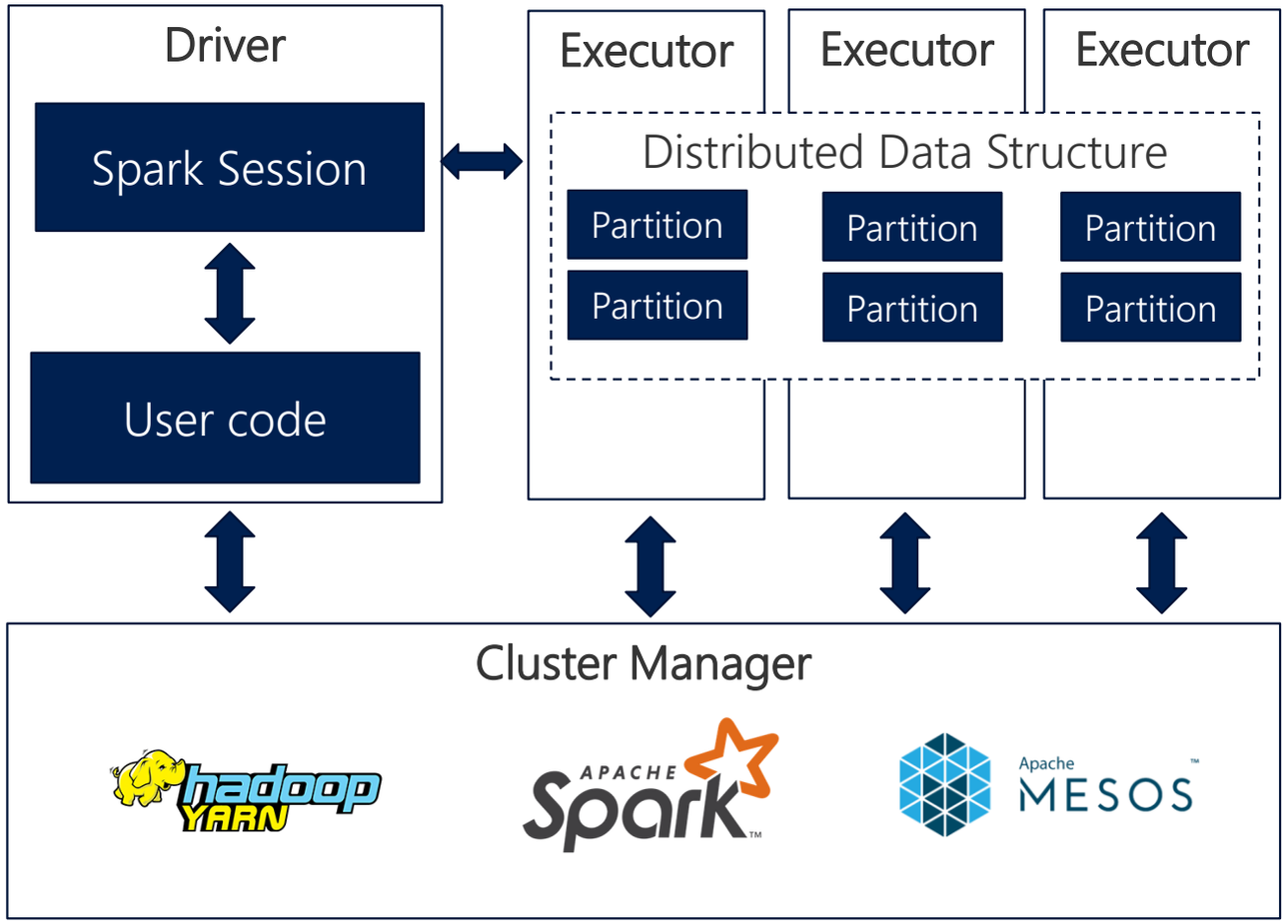
Data Frames with user objects and custom code



Transformations and Actions



Inside a Spark Application



Spark on HDInsight

Azure HDInsight

A Cloud Spark and Hadoop service for the Enterprise

- Reliable with an industry leading SLA
- Enterprise-grade security and monitoring
- Productive platform for developers and scientists
- Cost effective cloud scale
- Integration with ISV applications
- Easy for administrators to manage
- 63% lower TCO than deploy your own Hadoop on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

Spark on Azure HDInsight

Fully Managed Service

Provision cluster with a click of a mouse
100% open source Apache Spark and Hadoop bits
Latest releases of Spark (1.6.3 and 2.1.1 are latest supported releases)
Fully supported by Microsoft and Hortonworks
99.9% Azure Cloud SLA
Certifications: PCI, ISO 27018, SOC, HIPAA, EU-MC

Optimized for data exploration, experimentation & development

Jupyter/Zeppelin Notebooks (scala, python, automatic data visualizations)
IntelliJ/Eclipse plugins (job submission, remote debugging)
ODBC connector for Power BI, Tableau, Qlik, SAP, Excel, etc

R Server on HDInsight

Spark cluster

… with a Microsoft R Server edge node

… R installed across the nodes

… enterprise-scale R analytics

… multi-threaded math libraries

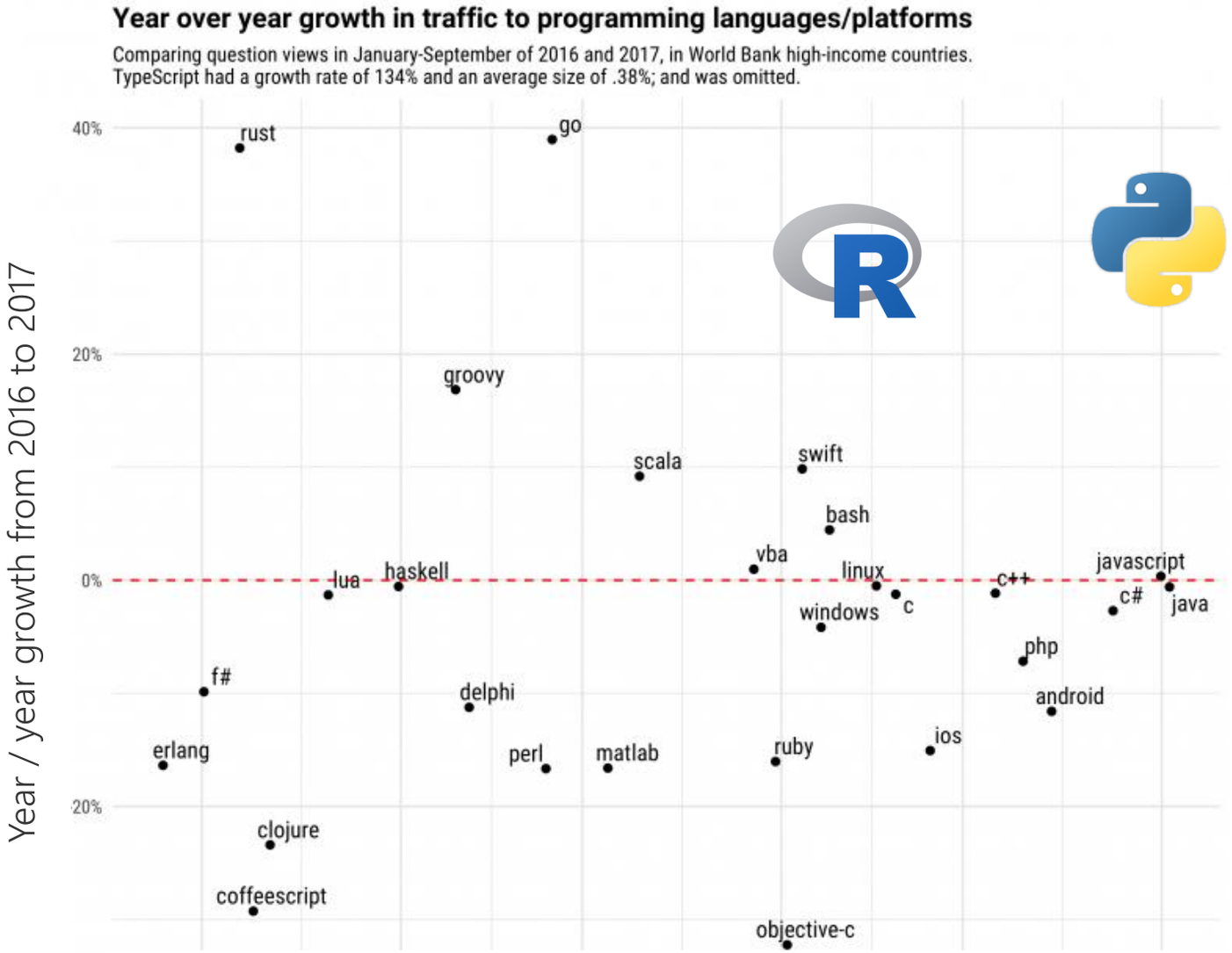
Demo

Hello, HDInsight

SparkR

R

- Statistical programming language
- Rich ecosystem of packages in CRAN
  - 10,000+ packages
- Powerful data visualization libraries
- Created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
  - Initial version in 1995. Based on S language (1975)
- Interpreted and **single threaded**

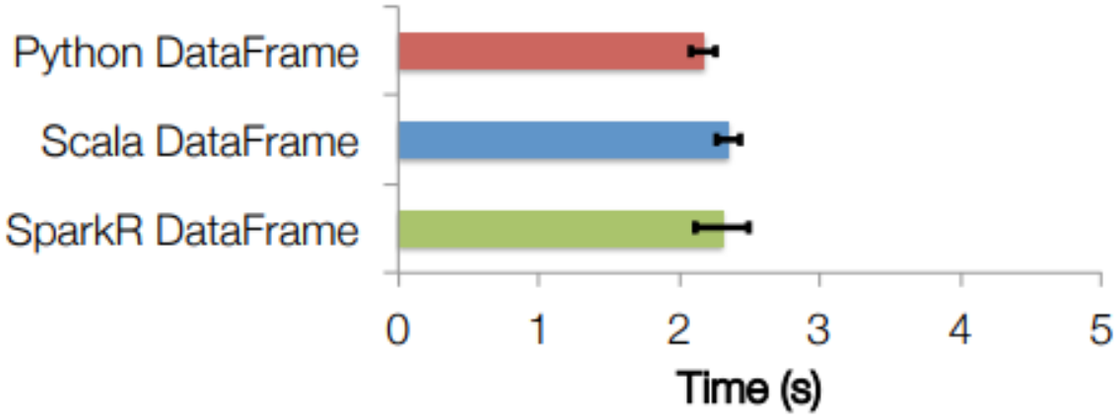


Average % of Stack Overflow visits across the two years (log scale)

Source: The Impressive Growth of R by David Robinson, https://stackoverflow.blog/2017/10/10/impressive-growth-r/

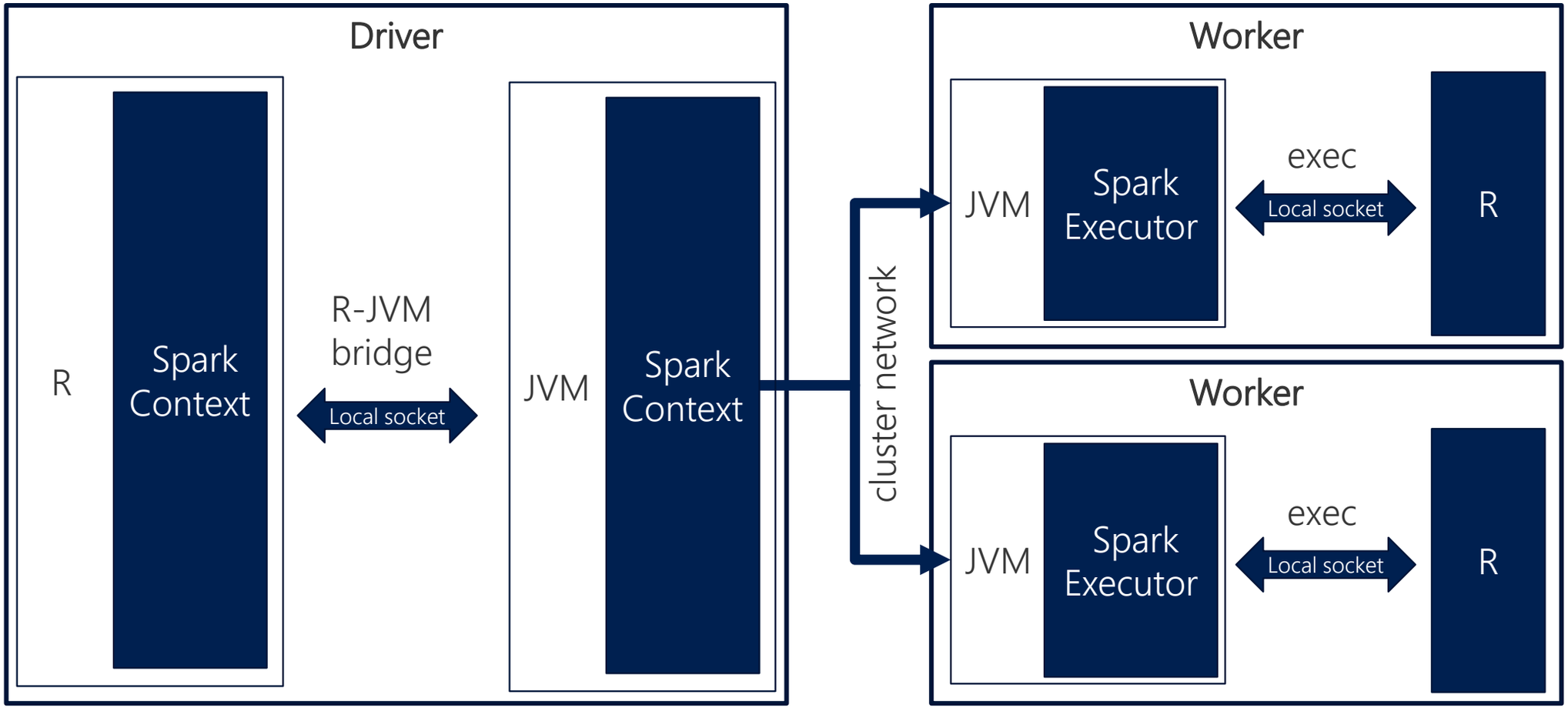
SparkR

- Light-weight R frontend to Spark
- Exposes Spark's DataFrame API
- MLlib bindings
- R package
- REPL (`sparkR`)



Source: http://people.csail.mit.edu/matei/papers/2016/sigmod_sparkr.pdf

SparkR Architecture



Source: http://people.csail.mit.edu/matei/papers/2016/sigmod_sparkr.pdf

Creating a SparkSession

```
# Load SparkR library
library(SparkR, lib.loc = “/path/to/package”)
# Start a spark session
sparkR.session()
# Inspect spark session
sparkR.conf()
# Stop a spark session
sparkR.session.stop()
```

Creating a SparkSession

```
# Start a spark session w/ additional config
sparkR.session(master = "yarn",
     sparkConfig = list('spark.executor.memory' = '20g',
                        'spark.executor.instances' = '20',
                        'spark.executor.cores' = '4',
                        'spark.driver.memory' = '4g'))
```

Reading and writing data

```
read.json()
read.orc()
read.parquet()
read.text()
read.jdbc()
read.df()

# corresponding write.format() methods
```

More at spark-packages.org

SparkPackages

Feedback Register a package Login Find a package

A community index of third-party packages for Apache Spark.

Showing packages 1 - 50 out of 392

Next >

All (392) Core (14) Data Sources (48) Machine Learning (79) Streaming (54) Graph (18) PySpark (17) Applications (14) Deployment (12) Examples (24) Tools (30)

spark-als

Another, hopefully better, implementation of ALS on Spark (already merged into MLlib)

@mengxr / Latest release: 0.1.0 (2014-11-27) / BSD 3-Clause / (1)

2 ml 1 mllib 1 recommendation

mllib-grid-search

An example project for doing grid search in MLlib

@spark-ml / Latest release: 0.0.1 (2014-11-27) / BSD 3-Clause / (2)

1 ml 1 example 1 examples

spark-avro

Integration utilities for using Spark with Apache Avro data

@databricks / Latest release: 4.0.0-s_2.11 (2017-10-30) / Apache-2.0 / (12)

6 sql 4 input 4 avro

Converting between R and Spark

```
# R DataFrame → Spark DataFrame
sdf <- createDataFrame(rdf)

# Spark DataFrame → R DataFrame
rdf <- collect(sdf)
```

SparkSQL in SparkR

```
# Query a hive table
sdf <- sql("SELECT * FROM hiveTable")

# On an existing Spark DataFrame
createOrReplaceTempView(sdf, "myTable")
sdf2 <- sql("SELECT * FROM myTable")
```

Manipulating Data in SparkR

```
# Transformations
filter()
select()
join()
groupBy()
pivot()
explode()
summarize()
sample()
distinct()
arrange()
# and more …
```

```
# Actions
collect()
head()
first()
nrow()
take()
```

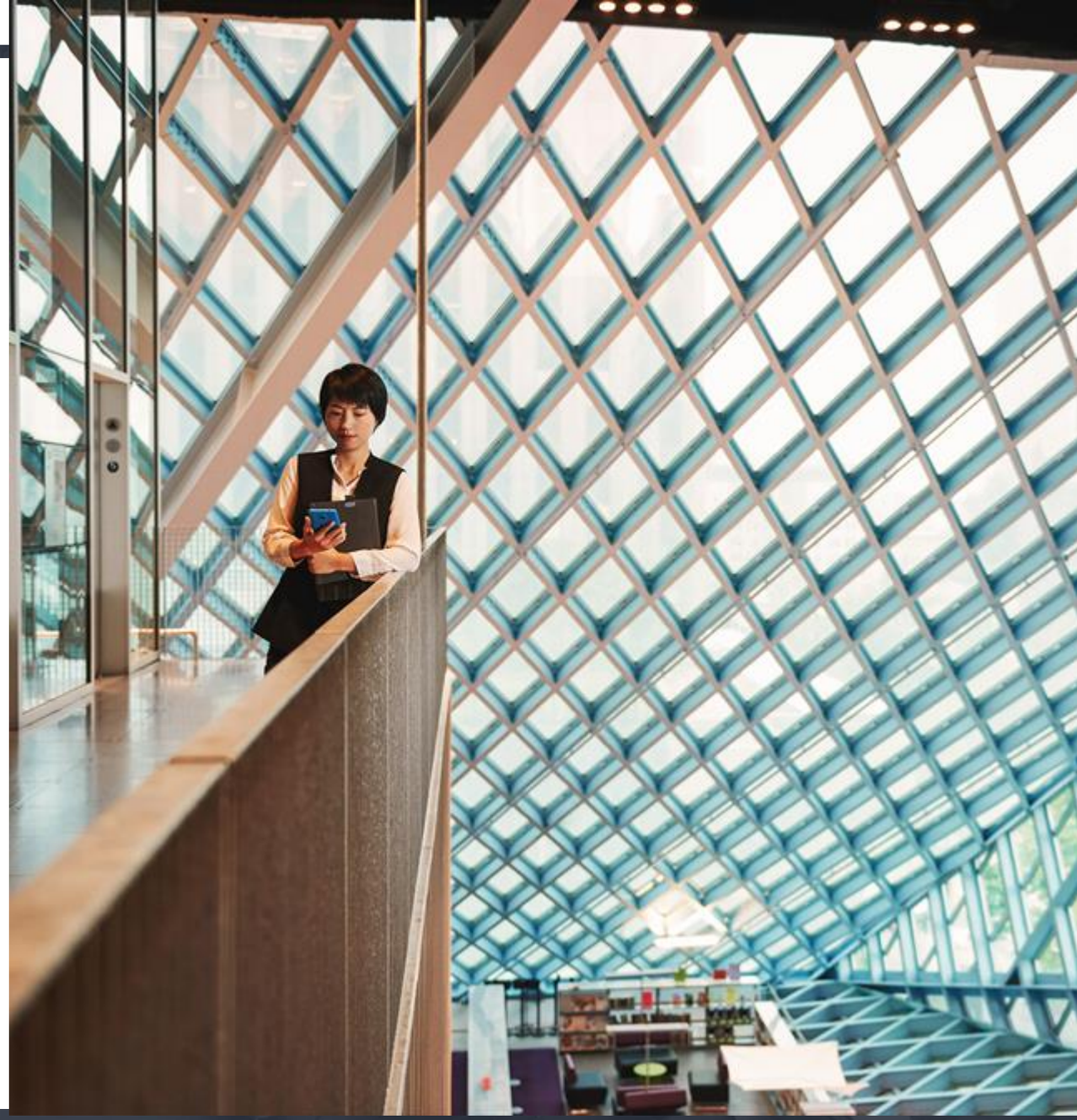
Seattle Public Library

Predicting daily borrowed book types and collections

Given:

- Checkout history
- Library collection

Data source:
https://data.seattle.gov



Demo

Hello, SparkR

Window Functions

- Operates on a group of rows while returning a single value per row
- Aggregate functions are **n to 1**, Window functions are **n to n**
- Great for rolling operations

Window Functions

```
# Create window spec
ws <- orderBy(windowPartitionBy("cyl"), "mpg")

# Apply rank() within window spec
sdfCarsRank <- withColumn(sdfCars, "rank",
    over(rank(), ws))
```

Demo

SparkR: Window Functions and Feature Engineering

Machine Learning w/ SparkR

Machine Learning models in SparkR

Difinity

Classification

- Logistic Regression
- Multilayer Perceptron (MLP)
- Naive Bayes
- Linear Support Vector Machine *(Spark 2.2 only)*

Regression

- Accelerated Failure Time (AFT) Survival Model
- Generalized Linear Model (GLM)
- Isotonic Regression

Tree

- Gradient Boosted Trees for Regression and Classification
- Random Forest for Regression and Classification

Clustering

- Bisecting k-means *(Spark 2.2 only)*
- Gaussian Mixture Model (GMM)
- K-Means
- Latent Dirichlet Allocation (LDA)

Collaborative Filtering

- Alternating Least Squares (ALS)
- Frequent Pattern Mining *(Spark 2.2 only)*
- FP-growth *(Spark 2.2 only)*

Statistics

- Kolmogorov-Smirnov Test

Split data

```
# Split into Train and Test
sdfSplit <- sdfCheckoutsWeekly %>%
    randomSplit(weights = c(7, 3), seed = 123)
sdfTrain <- sdfSplit[[1]]
sdfTest <- sdfSplit[[2]]
```

Example: Random Forest

```
# Fit model
model <- spark.randomForest(train,
    label ~ feature1 + feature2, type =
    "regression",   maxDepth = 5,   maxBins = 16,
    numTrees = 20,  seed = 10)

# Predictions
predictions <- predict(model, newdata = test)
```

Example: Random Forest

```
# Fit model
model <- spark.randomForest(train,
    label ~ feature1 + feature2, type =
    "regression",   maxDepth = 5, maxBins = 16,
    numTrees = 20,  seed = 10)

# Predictions
predictions <- predict(model, newdata = test)
```

Model Persistence in SparkR

```
# Save model
write.ml(model, path = “/model/path”)

# Load model
model <- read.ml(path = “/model/path”)
```

Demo

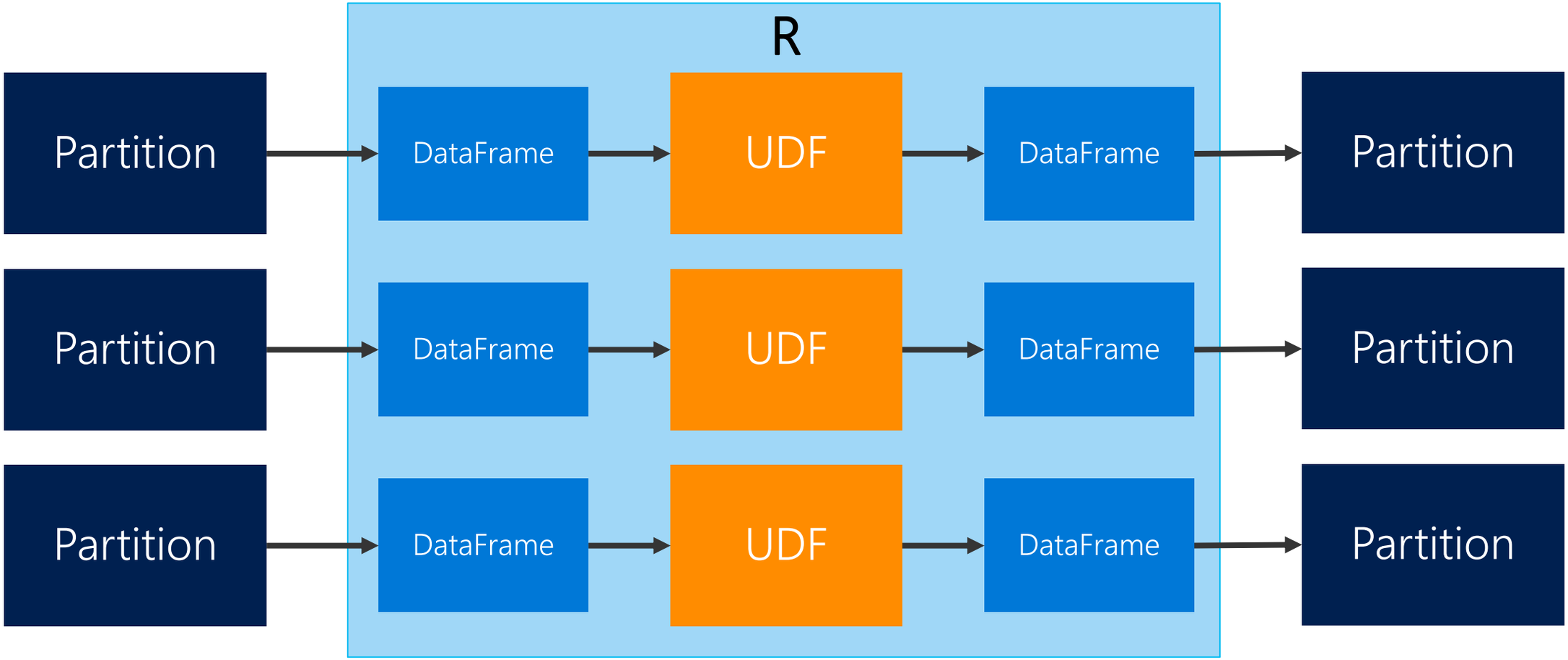
SparkR: Machine Learning

Parallelizing Native R

User Defined Functions (UDFs)

- By partition – `dapply` or `dapplyCollect`
- By group – `gapply` or `gapplyCollect`
- Distributed R – `lapply`

Parallel Processing By Partition (dapply)



Source: Scalable Data Science with SparkR: Spark Summit East talk by Felix Cheung

Example: dapply

```
# Select specific columns
sdfSubCars <- select(sdfCars, "model", "mpg")
# Define schema
schema <- "model STRING, mpg DOUBLE, kmpg DOUBLE"
# Use dapply
out <- dapply(sdfSubCars, function(x) {
     x <- cbind(x, x$mpg * 1.61)
}, schema)
```

Demo

Installing R packages using script actions

Challenges with SparkR

- Can be hard to debug ML jobs
- Not as mature as the Python, Java, Scala interfaces
- Can't call directly feature transformers/extractors
  - Currently, fixed pipelines
  - Not all Spark MLlib functions are directly exposed/ported
- Collects can be slow

More to checkout…

R on Spark

- ScaleR / RxSpark (Machine Learning Server / R Server)
- Sparklyr
- H2O R

Managed Spark on Azure

- Azure Databricks

Q & A

Microsoft Difinity

Thank you!

Lace Lofranco
Senior Software Development Engineer
Microsoft
lace.lofranco@microsoft.com

Thanks to our sponsors

Platinum Sponsors

Microsoft

RADACAD

Exhibitors

altis

redgate

TIMEXTENDER

Media and Venue sponsors

CORDIS

STAGING CONNECTIONS