ggplot2.SparkR: Rebooting ggplot2 for Scalable Big Data Visualization

Jonghyun Bae[†], Sangoh Jeong[*], Wenjing Jin[†] and Jae W. Lee[†]

[†]Sungkyunkwan University

[*]SK Telecom

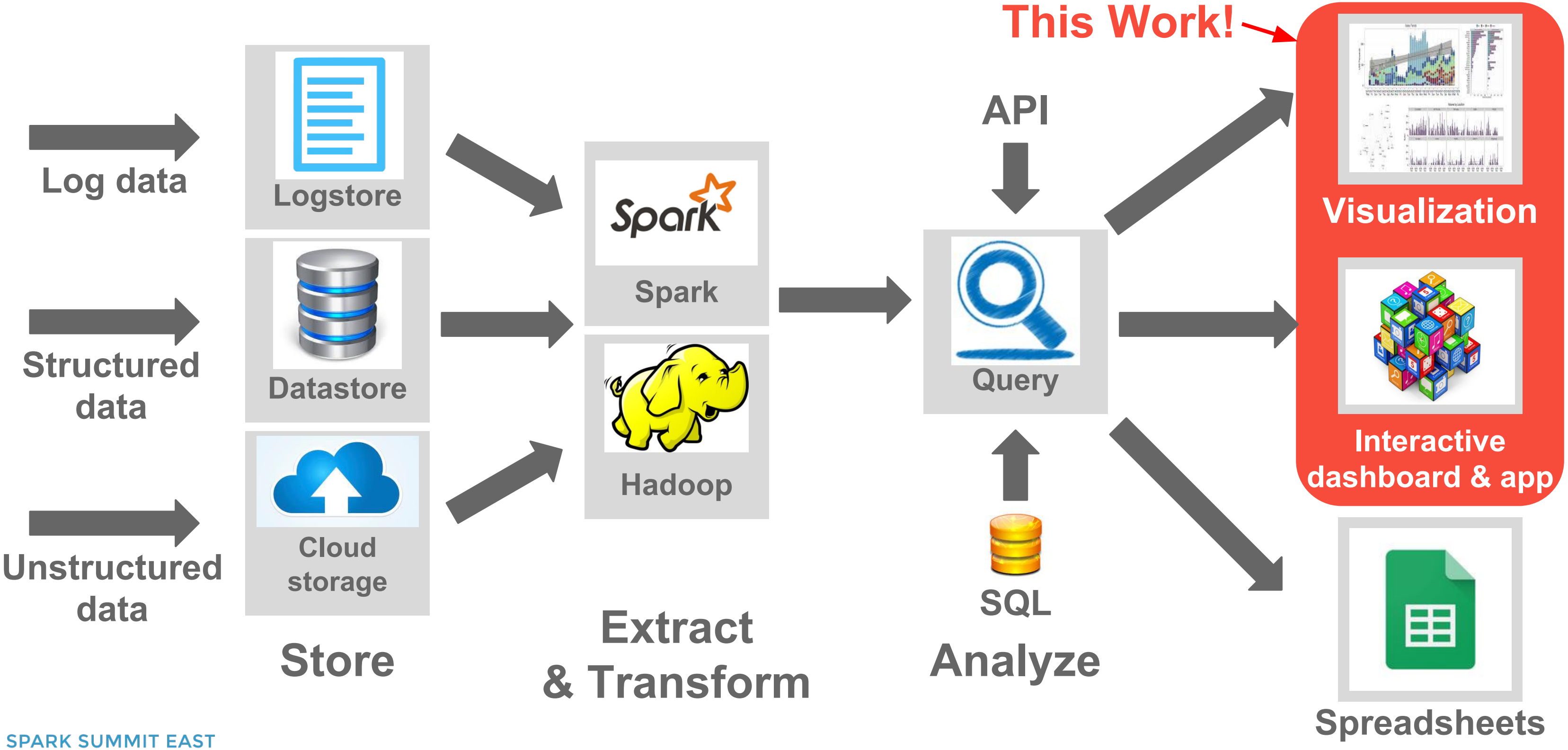
SPARK SUMMIT EAST
DATA SCIENCE AND ENGINEERING AT SCALE
FEBRUARY 16-18, 2016 NEW YORK CITY

Speakers

- Sangoh Jeong (sangoh.jeong@sk.com)
  - Senior Manager at SK telecom in Korea
  - Interested in R Spark Machine Learning Spark MLlib

- Jonghyun Bae (jonghbae@skku.edu)
  - Graduate student at SUNGKYUNKWAN UNIVERSITY 1398
  - Interested in R, JavaScript and Spark Google SUMMER of CODE TENTH YEAR 2014

Big Data Analytics Pipeline



Log data

Structured data

Unstructured data

Logstore

Datastore

Cloud storage

Store

Spark

Hadoop

Extract & Transform

API

Query

SQL

Analyze

This Work!

Visualization

Interactive dashboard & app

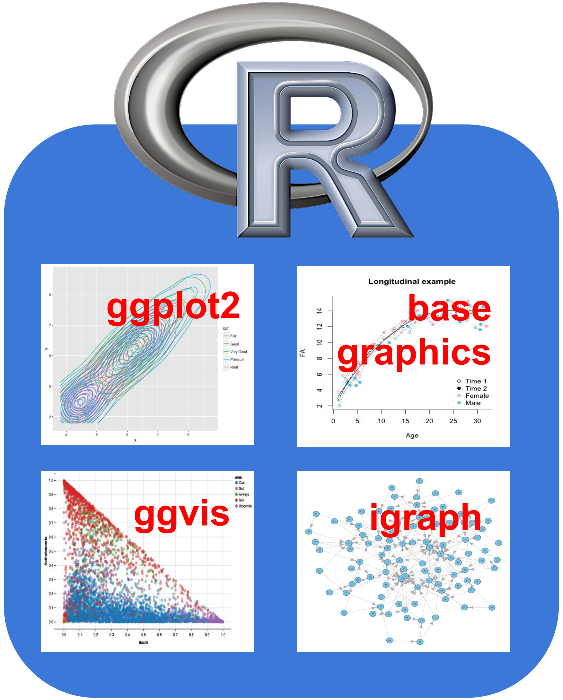
Spreadsheets

Source: http://www.slideshare.net/BigDataCloud/big-data-analytics-with-google-cloud-platform

Why Big Data Visualization?

- Case of a business unit at SK Telecom
  - Typical DB size: 70M records with 330 columns
  - Analyzes the DB using R on a single-node scale-up server
  - Has much bigger DBs that cannot be handled by this server
- The business unit’s visualization needs
  - Use of R
  - Easy-to-use APIs
  - Scalable solution for the bigger DBs

R Has Great Visualization Packages



But, these packages cannot process Spark DataFrames.

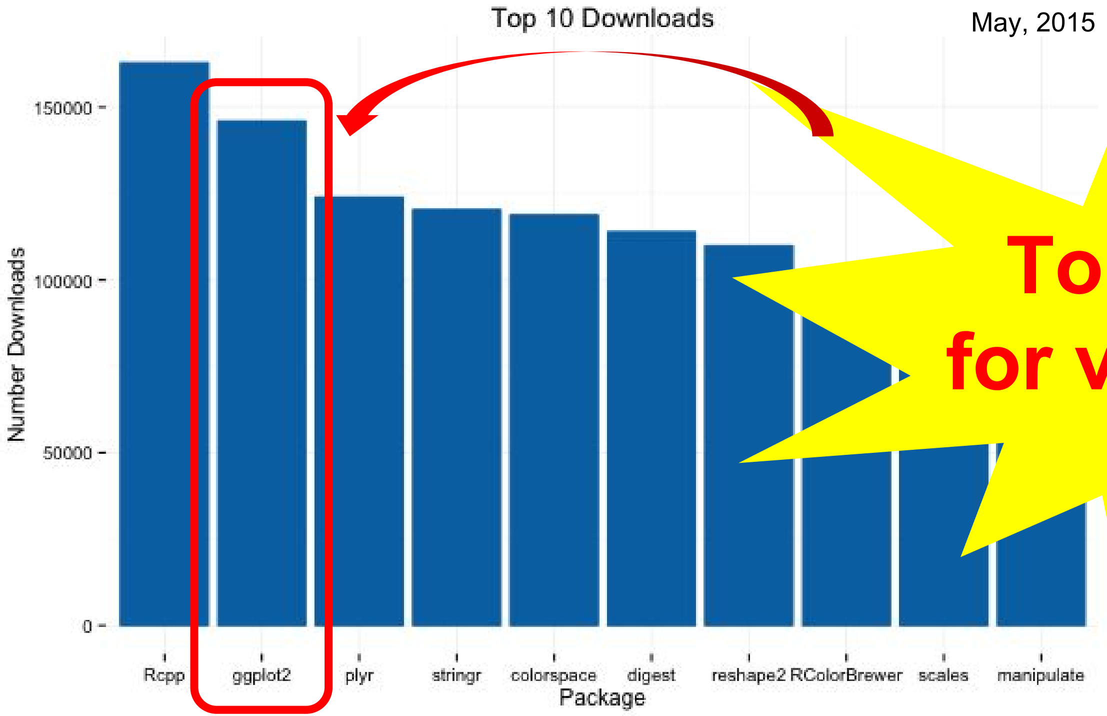
ggplot2

- (Arguably) the most popular visualization package for R
  - Based on the “layered” grammar of graphics
  - Making it easy to produce high-quality graphs
  - Limited to single node processing

“Base graphics are good for drawing pictures;
ggplot2 graphics are good for understanding the data”

(Hadley Wickham, Creator of ggplot2, 2012)

Top 10 packages in R

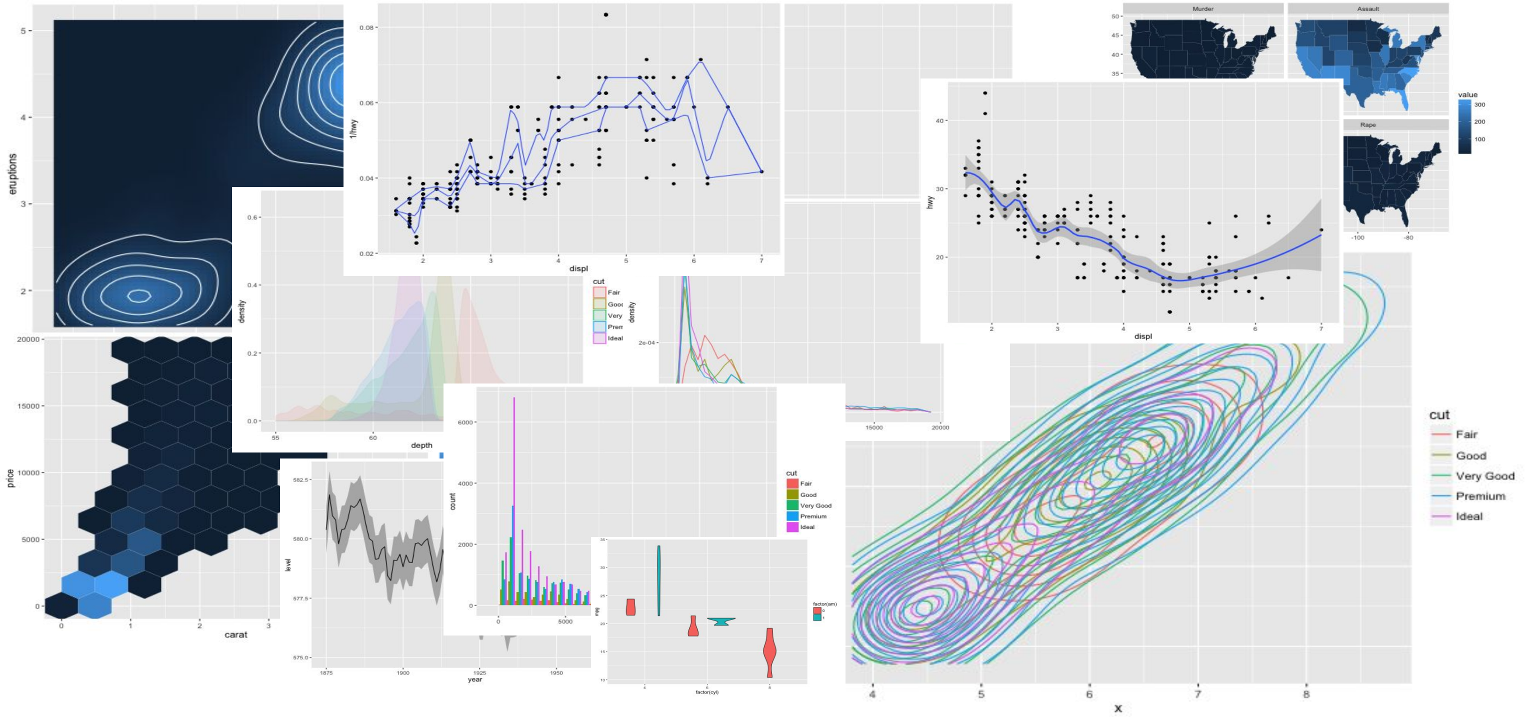


May, 2015

Top package for visualization

Source: http://www.computerworld.com/article/2920117/business-intelligence/most-downloaded-r-packages-last-month.html

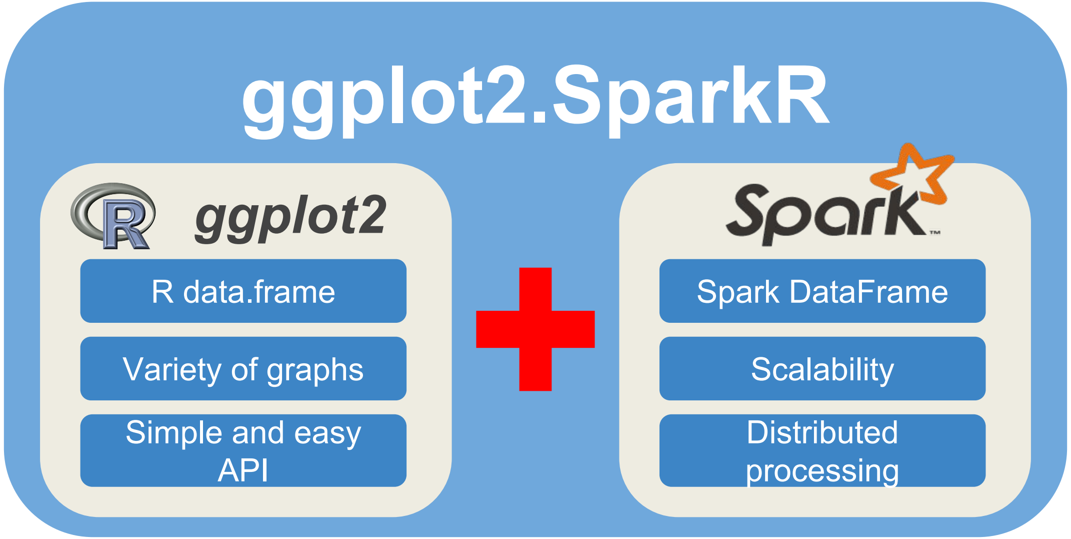
ggplot2: Example Plots



Source: ggplot2 documentation, http://docs.ggplot2.org/current

ggplot2.SparkR = SparkR+ggplot2!

- An R package extending ggplot2 to visualize big data represented in Spark DataFrame



ggplot2.SparkR

ggplot2

R data.frame

Variety of graphs

Simple and easy API

Spark

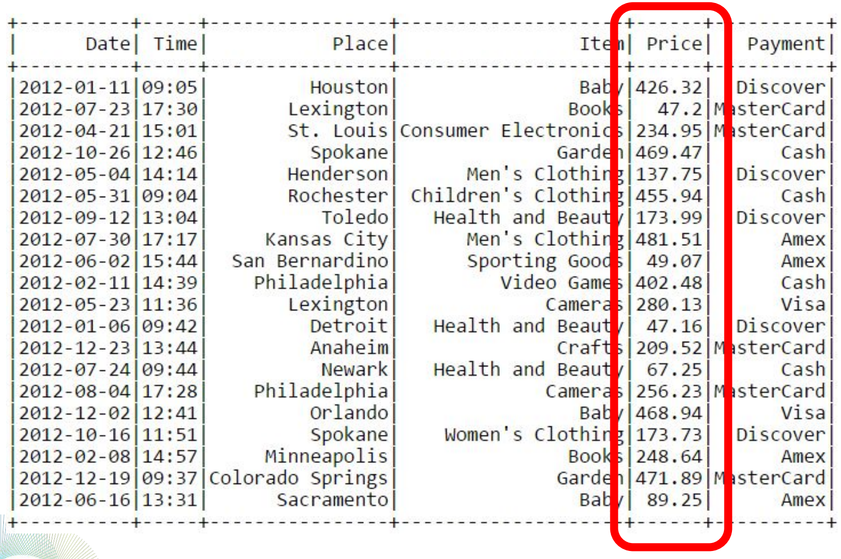
Spark DataFrame

Scalability

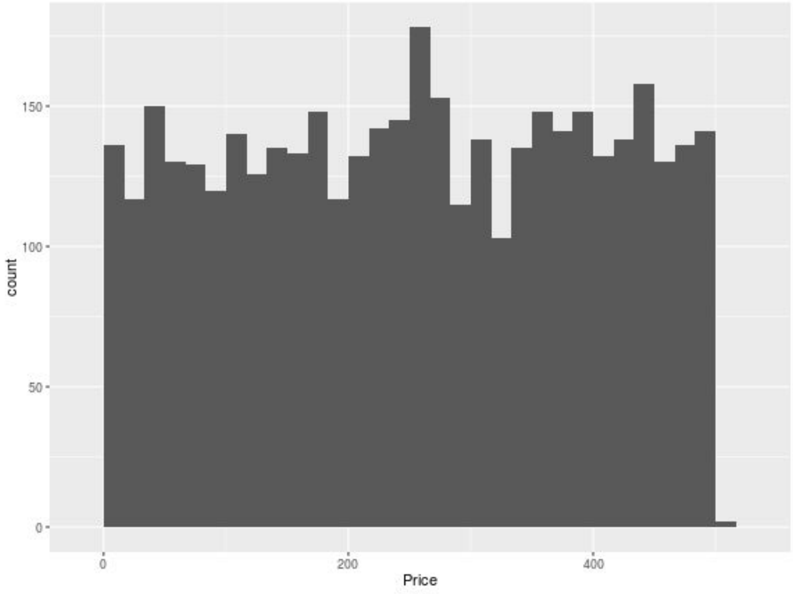
Distributed processing

ggplot2.SparkR Simplifies Plotting (1)

Example: Draw a histogram using DataFrame



```
+----------+-----+-----------------+--------------------+------+----------+
|      Date| Time|            Place|                Item| Price|   Payment|
+----------+-----+-----------------+--------------------+------+----------+
|2012-01-11|09:05|          Houston|                Baby|426.32|  Discover|
|2012-07-23|17:30|        Lexington|               Books|  47.2|MasterCard|
|2012-04-21|15:01|        St. Louis|Consumer Electronics|234.95|MasterCard|
|2012-10-26|12:46|          Spokane|              Garden|469.47|      Cash|
|2012-05-04|14:14|        Henderson|        Men's Clothing|137.75|  Discover|
|2012-05-31|09:04|        Rochester|  Children's Clothing|455.94|      Cash|
|2012-09-12|13:04|           Toledo|     Health and Beauty|173.99|  Discover|
|2012-07-30|17:17|      Kansas City|        Men's Clothing|481.51|      Amex|
|2012-06-02|15:44|   San Bernardino|       Sporting Goods| 49.07|      Amex|
|2012-02-11|14:39|     Philadelphia|          Video Games|402.48|      Cash|
|2012-05-23|11:36|        Lexington|             Cameras|280.13|      Visa|
|2012-01-06|09:42|          Detroit|     Health and Beauty| 47.16|  Discover|
|2012-12-23|13:44|          Anaheim|              Crafts|209.52|MasterCard|
|2012-07-24|09:44|           Newark|     Health and Beauty| 67.25|      Cash|
|2012-08-04|17:28|     Philadelphia|             Cameras|256.23|MasterCard|
|2012-12-02|12:41|          Orlando|                Baby|468.94|      Visa|
|2012-10-16|11:51|          Spokane|      Women's Clothing|173.73|  Discover|
|2012-02-08|14:57|      Minneapolis|               Books|248.64|      Amex|
|2012-12-19|09:37| Colorado Springs|              Garden|471.89|MasterCard|
|2012-06-16|13:31|       Sacramento|                Baby| 89.25|      Amex|
+----------+-----+-----------------+--------------------+------+----------+
```



ggplot2.SparkR Simplifies Plotting (2)

BEFORE

```
# Pre-processing Spark DataFrame using SparkR API

range <- select(df, min(df$Price), max(df$Price))

breaks <- fullseq(range, diff(range / binwidth))

left <- breaks[-length(breaks)]; right <- breaks[-1]

breaks_df <- createDataFrame(sqlContext, data.frame(left = left,
  right = right))

histogram_df <- join(df, breaks_df, df$Price >= breaks_df$left &
  df$Price < breaks_df$right, “inner”)

histogram_df <- count(groupBy(histogram_df, “left”, “right”))


# Draw histogram chart using ggplot2 API

ggplot(collect(histogram_df), aes(xmin = left, xmax = right, ymin
  = 0, ymax = count)) + geom_rect()
```

AFTER

```
# It just takes one line!

ggplot(df, aes(x = Price)) + geom_histogram()
```

ggplot2.SparkR: Features

- **Scalable**
  - Beyond the capacity of single node (*cf*. ggplot2)
  - Performance scales to the number of nodes
- **Easy to use**
  - No changes to ggplot2 API
  - No training required for existing ggplot2 users
- **Readily deployable**
  - No modifications required for Spark
  - Using SparkR API only

The Rest of This Talk

Overview

How to Use It?

Architecture

Performance

Status & Plan

Summary

How to Use It?

Using ggplot2.SparkR is as easy as 1-2-3!

1

Install

R

2

Create

3

Draw

```
devtools::install_github
(“SKKU-SKT/ggplot2.SparkR”)
```

1. Install from Github

```
df <- read.json(sqlContext,
“hdfs://localhost:9000/dataset”)
```

2. Create DataFrame

Note that **df** is a Spark DataFrame object (not R data.frame).

```
ggplot(df, aes(x = Item, fill = 
Payment)) + geom_bar() + coord_flip()
```

3. Draw it (using ggplot2 API)!

Demo

1

Install

2

Create

3

Draw

Demo: Data Set

- Schema: Sales record from a department store chain
  - Source: `http://content.udacity-data.com/course/hadoop/forum_data.tar.gz`

Date	Item	Payment	Place	Price	Time
2012-01-11	Baby	Discover	Houston	426.32	09:05
2012-07-23	Books	MasterCard	Lexington	47.20	17:30
2012-04-21	Consumer Electronics	MasterCard	St. Louis	234.95	15:01
2012-10-26	Garden	Cash	Spokane	469.47	12:46

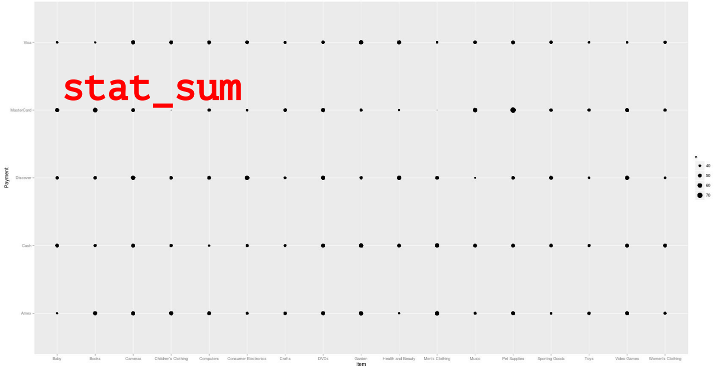
Supported Graph Types & Options

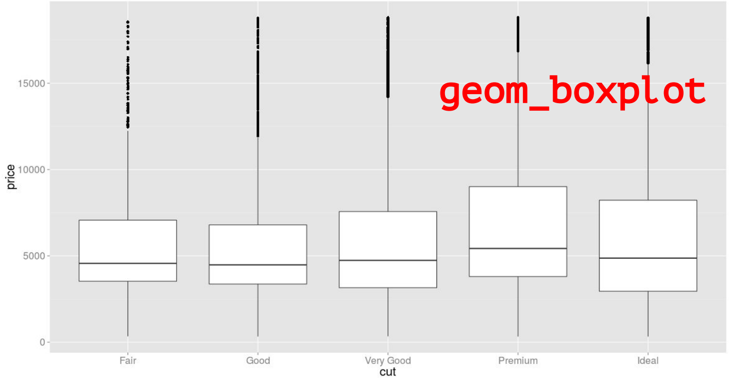
	Name	Descriptions
Graph types	geom_bar	Bars, rectangles with bases on x-axis.
	geom_histogram	Histogram.
	stat_sum	Sum unique values.
	geom_boxplot	Box and whiskers plot.
	geom_bin2d	Heatmap of 2d bin counts.
	geom_freqpoly	Frequency polygon.

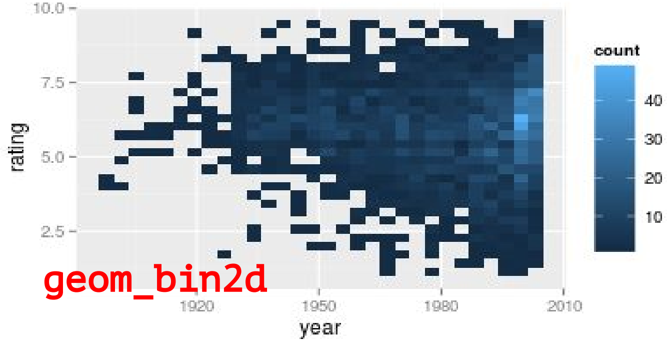
	Name	Descriptions
Positions	position_stack	Stack overlapping objects on top of one another
	position_fill	Same as above, but the range is standardized.
	position_dodge	Adjust position by dodging overlaps to the side
Facets	facet_null	Facet specification: a single panel
	facet_grid	Lay out panels in a grid
	facet_wrap	Wrap a 1d ribbon of panels into 2d
Scales	scale_x_log10	Put x-axis on a log scale
	scale_y_log10	Put y-axis on a log scale

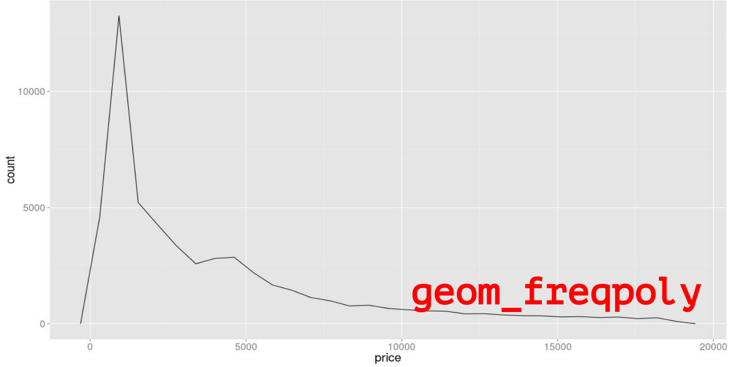
	Name	Description
Coords	coord_cartesian	Cartesian coordinates
	coord_flip	Flip cartesian coordinates
Ranges	xlim	Set the ranges of the x axis
	ylim	Set the ranges of the y axis
Texts	xlab	Change the label of x-axis
	ylab	Change the label of y-axis
	ggtitle	Change the graph title

Supported Graph Types

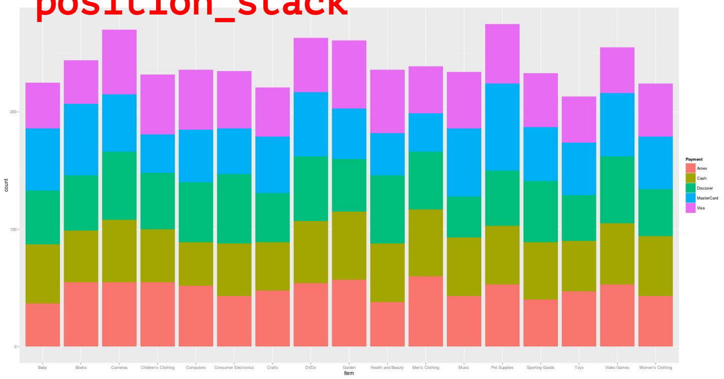

stat_sum

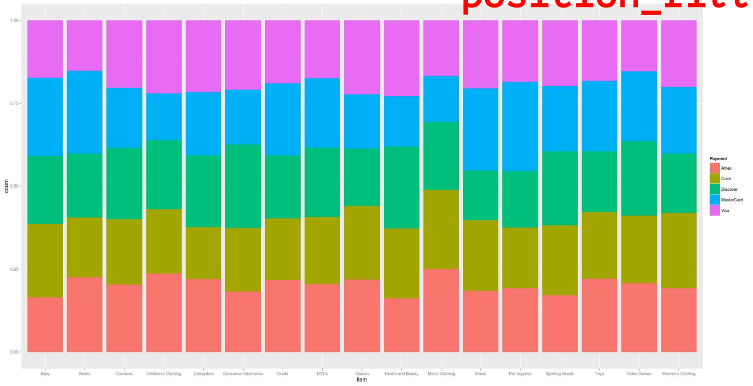

geom_boxplot

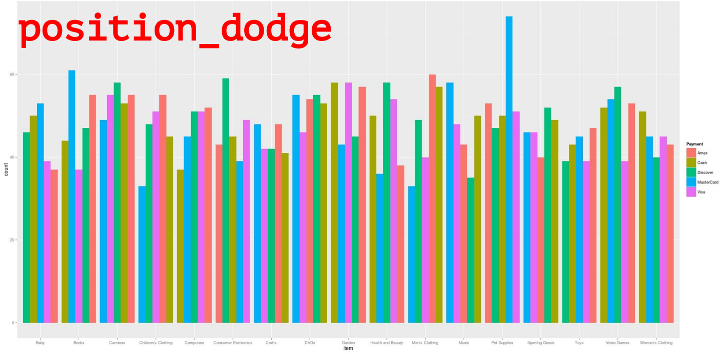

geom_bin2d

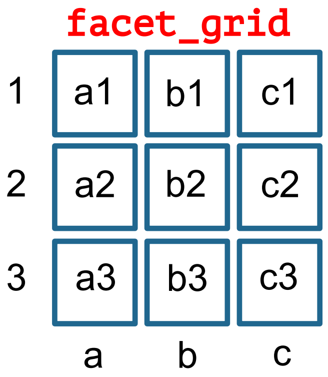

geom_freqpoly

Supported Graph Options

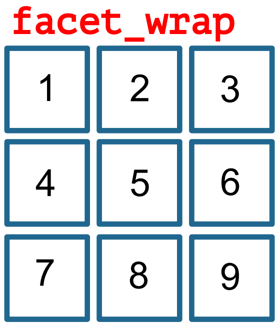

position_stack


position_fill


position_dodge


facet_grid

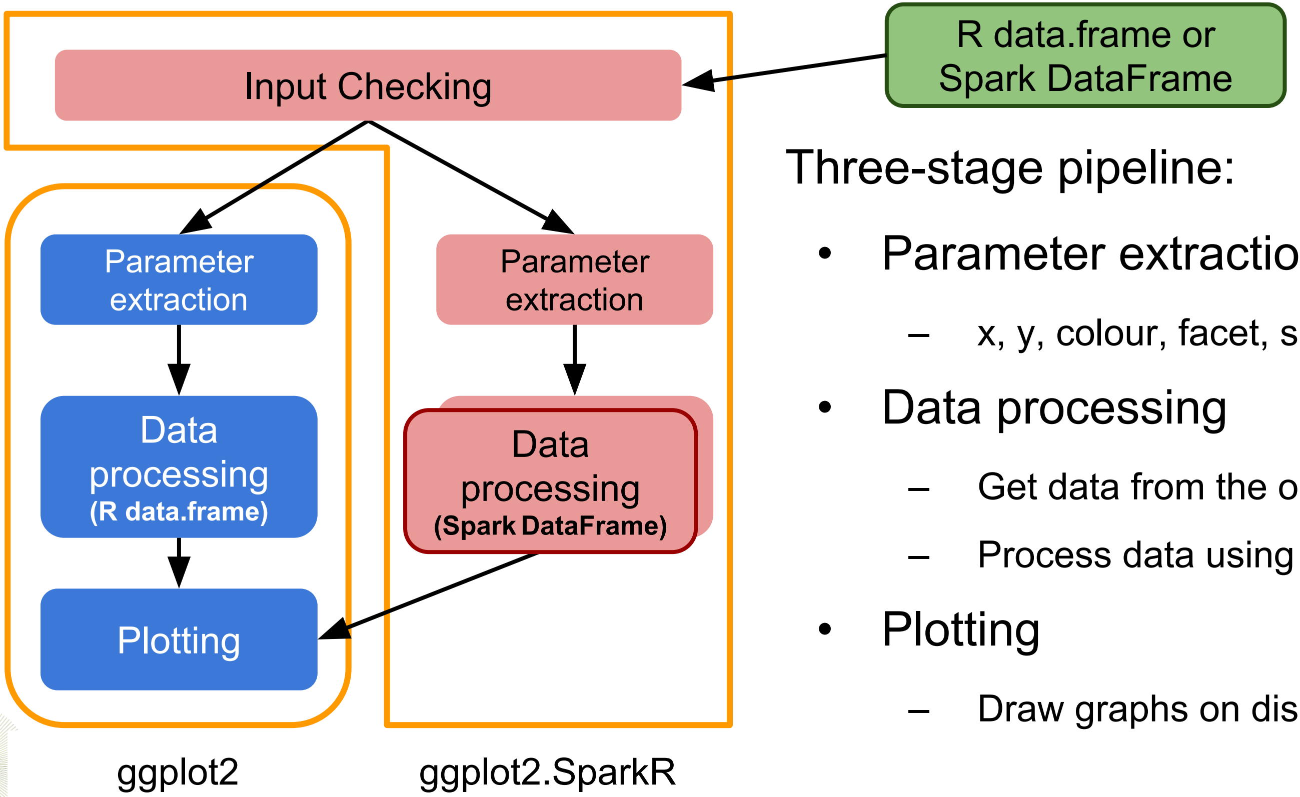
1	a1	b1	c1
2	a2	b2	c2
3	a3	b3	c3
	a	b	c


facet_wrap

1	2	3
4	5	6
7	8	9

Architecture

ggplot2.SparkR: Architecture (1)

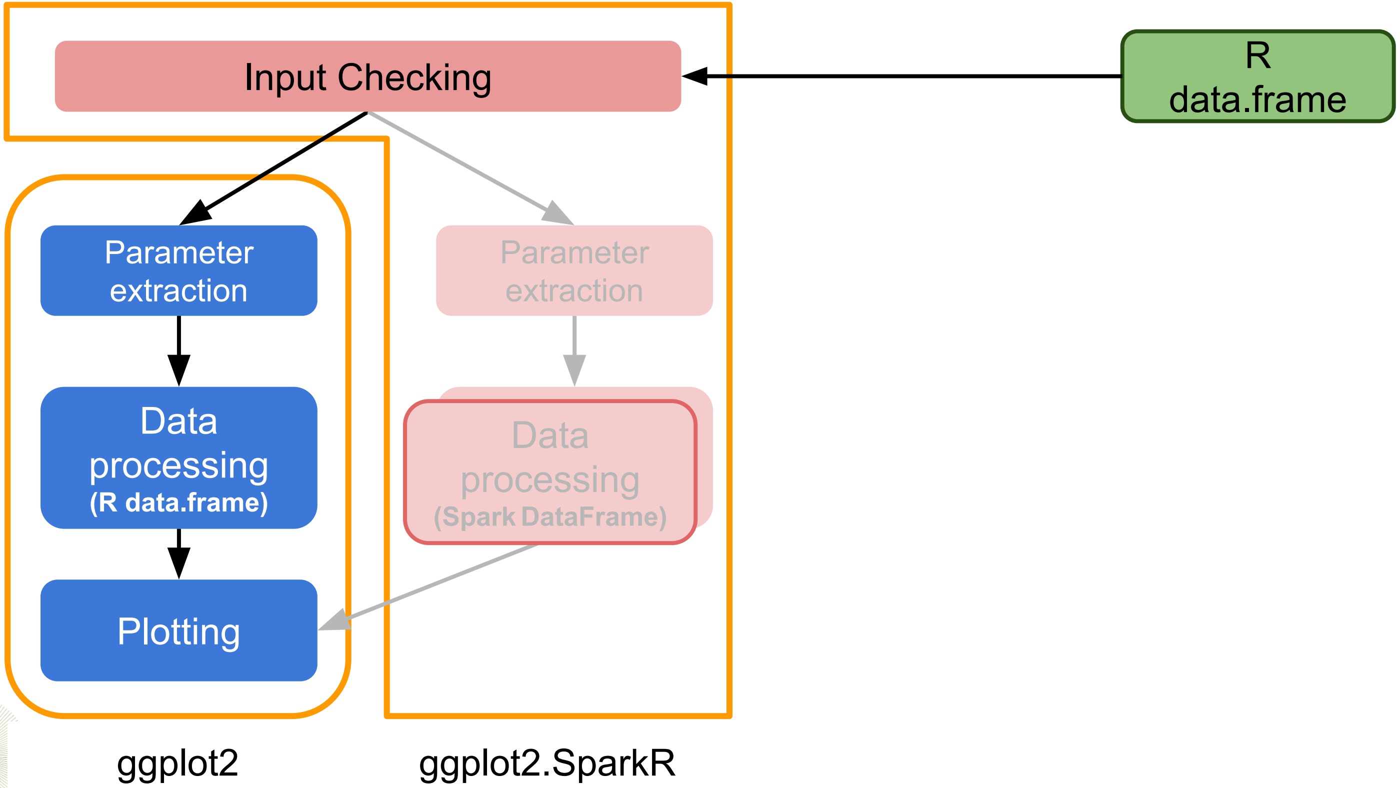


R data.frame or Spark DataFrame
Input Checking
Parameter extraction
Data processing (R data.frame)
Plotting
Parameter extraction
Data processing (Spark DataFrame)
ggplot2
ggplot2.SparkR

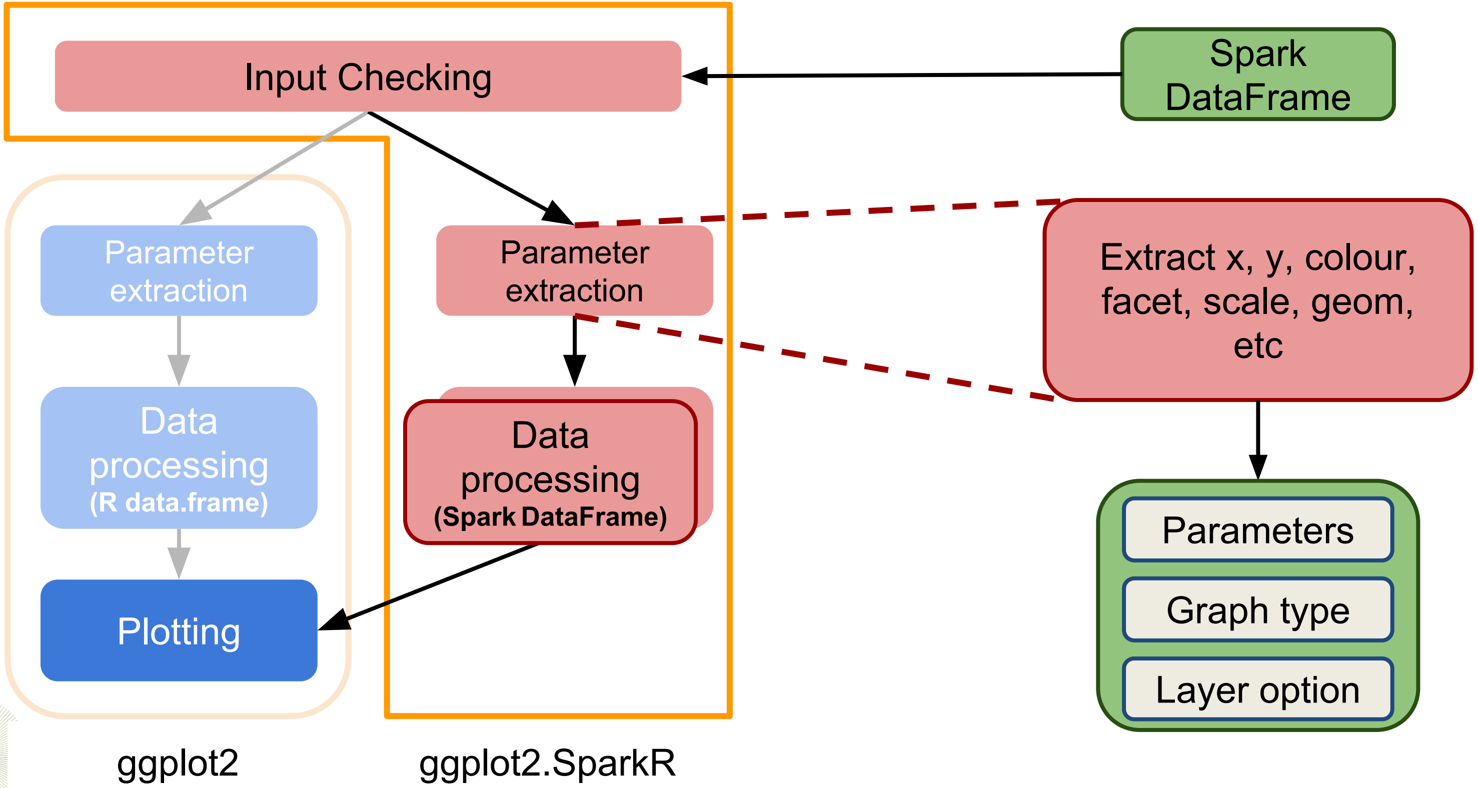
Three-stage pipeline:

- Parameter extraction
  - x, y, colour, facet, scale, geom, etc.
- Data processing
  - Get data from the original source
  - Process data using graph parameters
- Plotting
  - Draw graphs on display windows

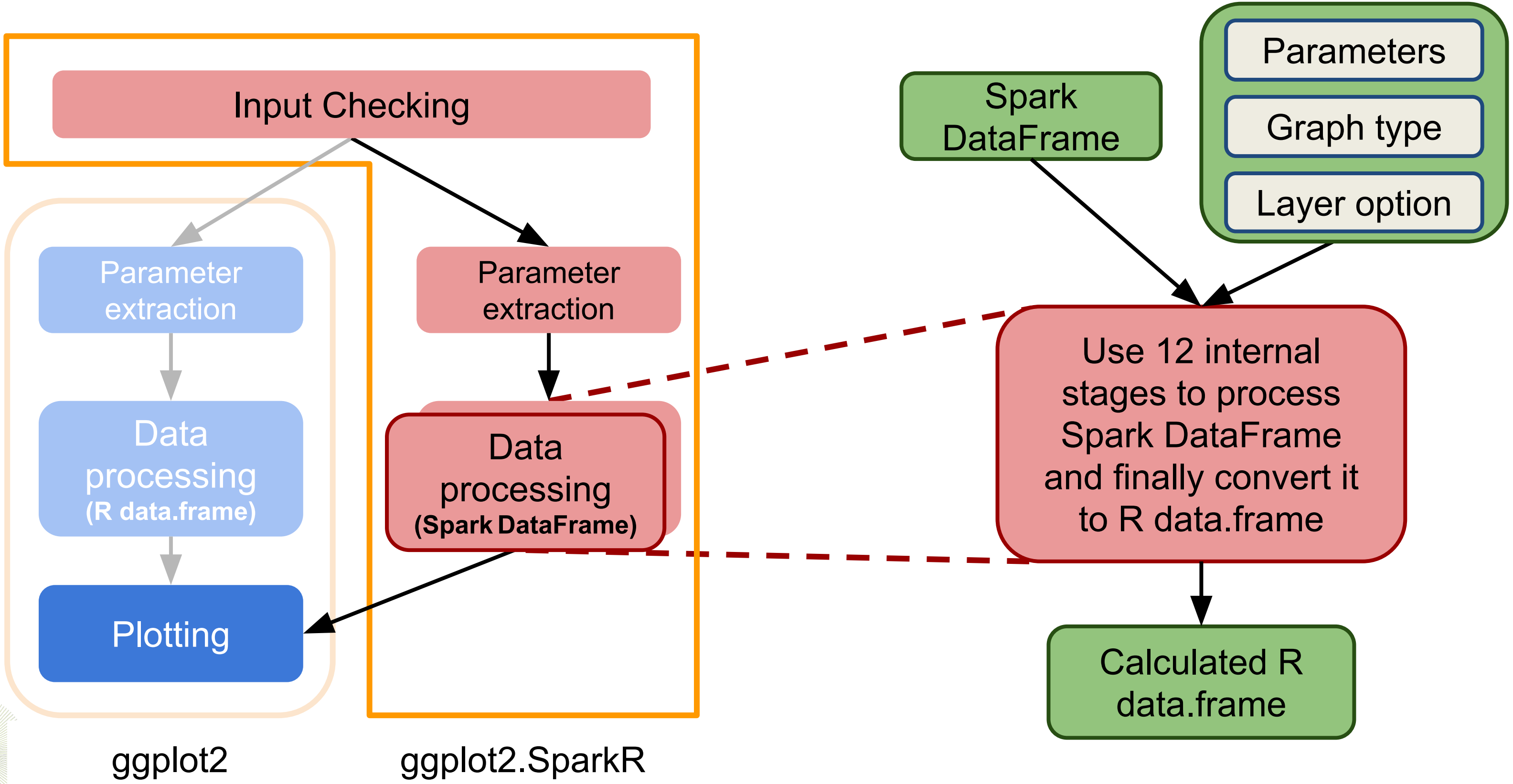
ggplot2.SparkR: Architecture (2)



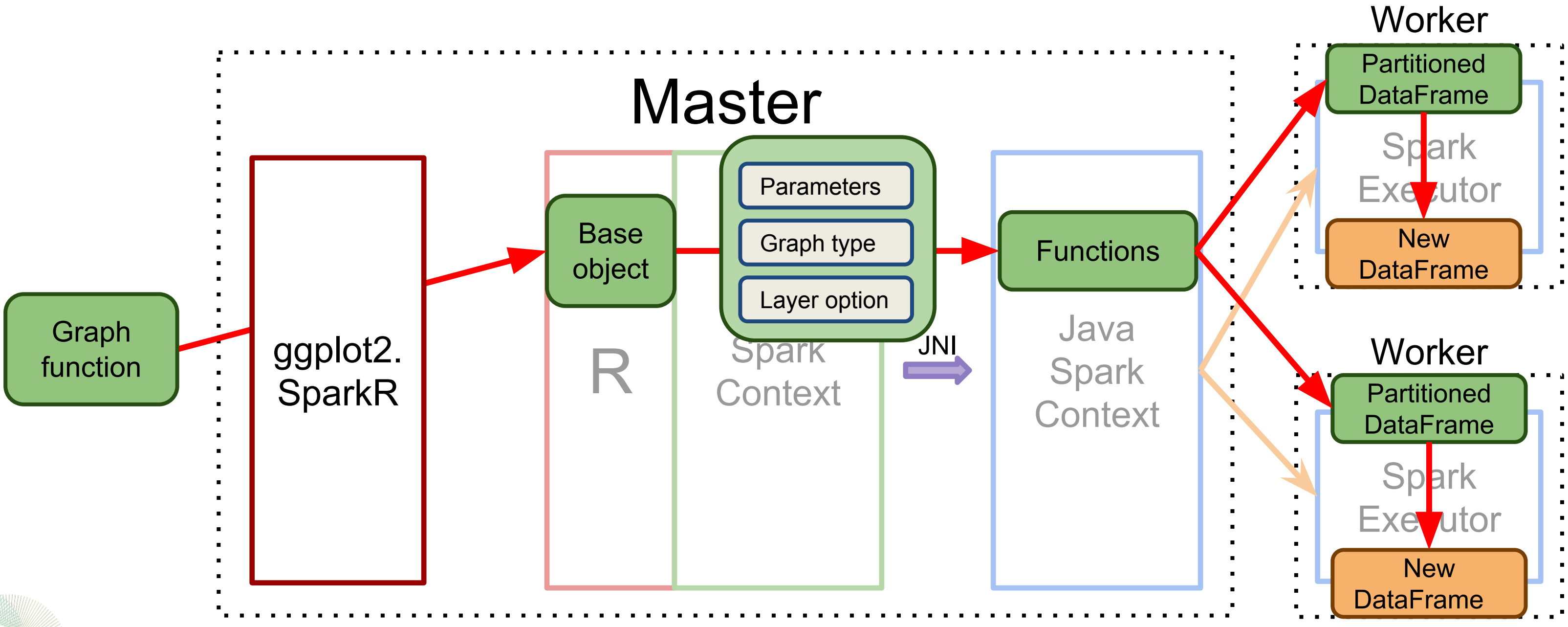
ggplot2.SparkR: Architecture (3)



ggplot2.SparkR: Architecture (4)

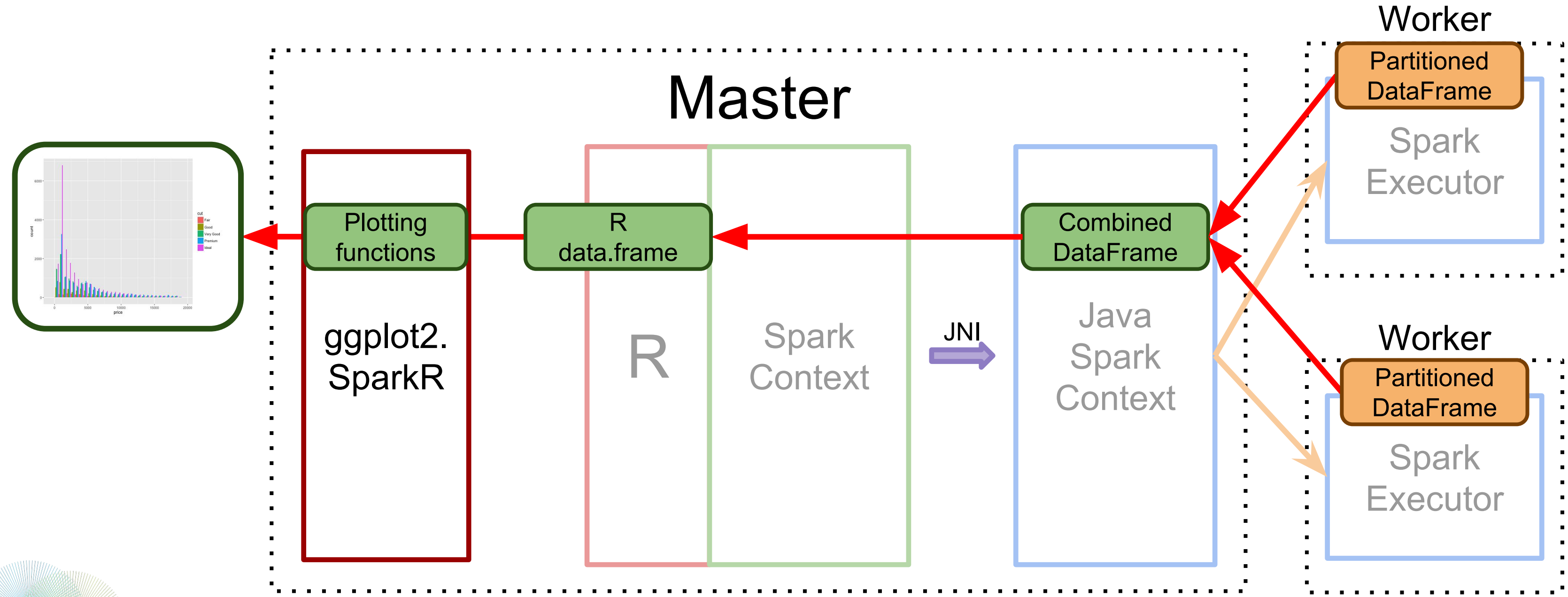


ggplot2.SparkR: Data Flow (1)



Source: https://spark-summit.org/2014/wp-content/uploads/2014/07/SparkR-SparkSummit.pdf

ggplot2.SparkR: Data Flow (2)



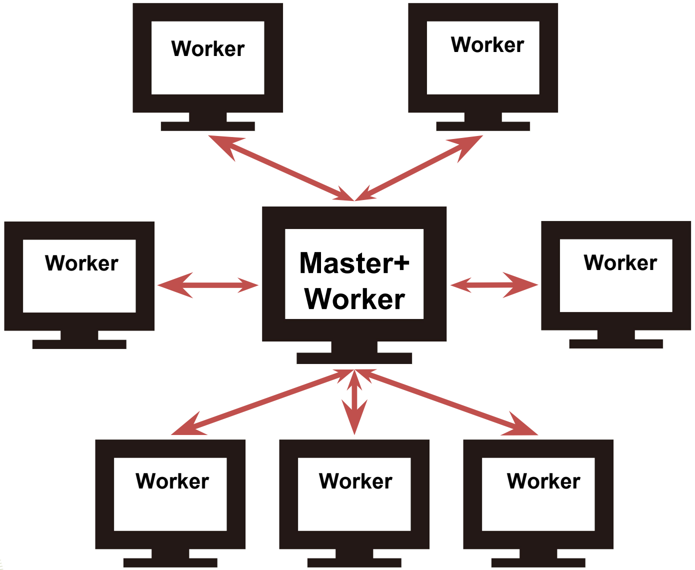
Source: https://spark-summit.org/2014/wp-content/uploads/2014/07/SparkR-SparkSummit.pdf

Performance

Experimental Setup (1)

- Cluster setup:

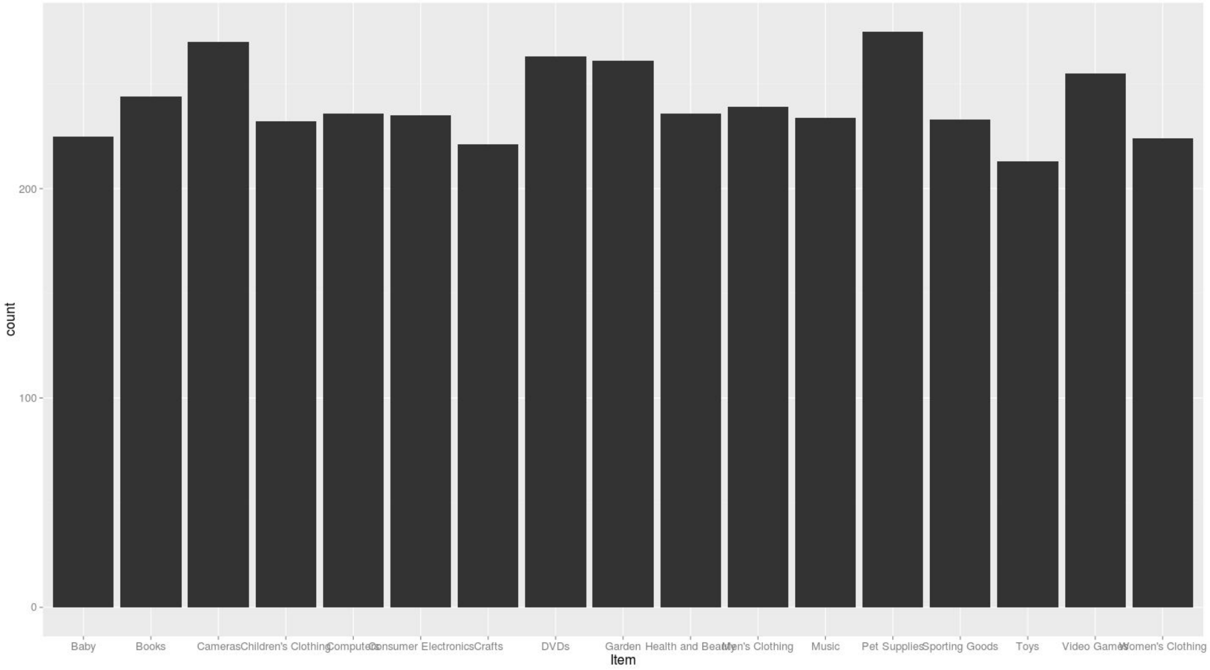
8-node Spark Cluster



Node Parameters	
CPU	Intel® i7-4790 (Haswell) 4GHz 8 cores
Memory	32GB DDR3 1600MHz
OS	Ubuntu 14.04 LTS
Hadoop	Ver. 1.2.1 (stable)
Spark	Ver. 1.5.0
R	Ver. 3.2.2
JDK	Ver. 1.8.0_60
Spark Worker	8 cores + 30GB / Worker
Network	Gigabit Ethernet

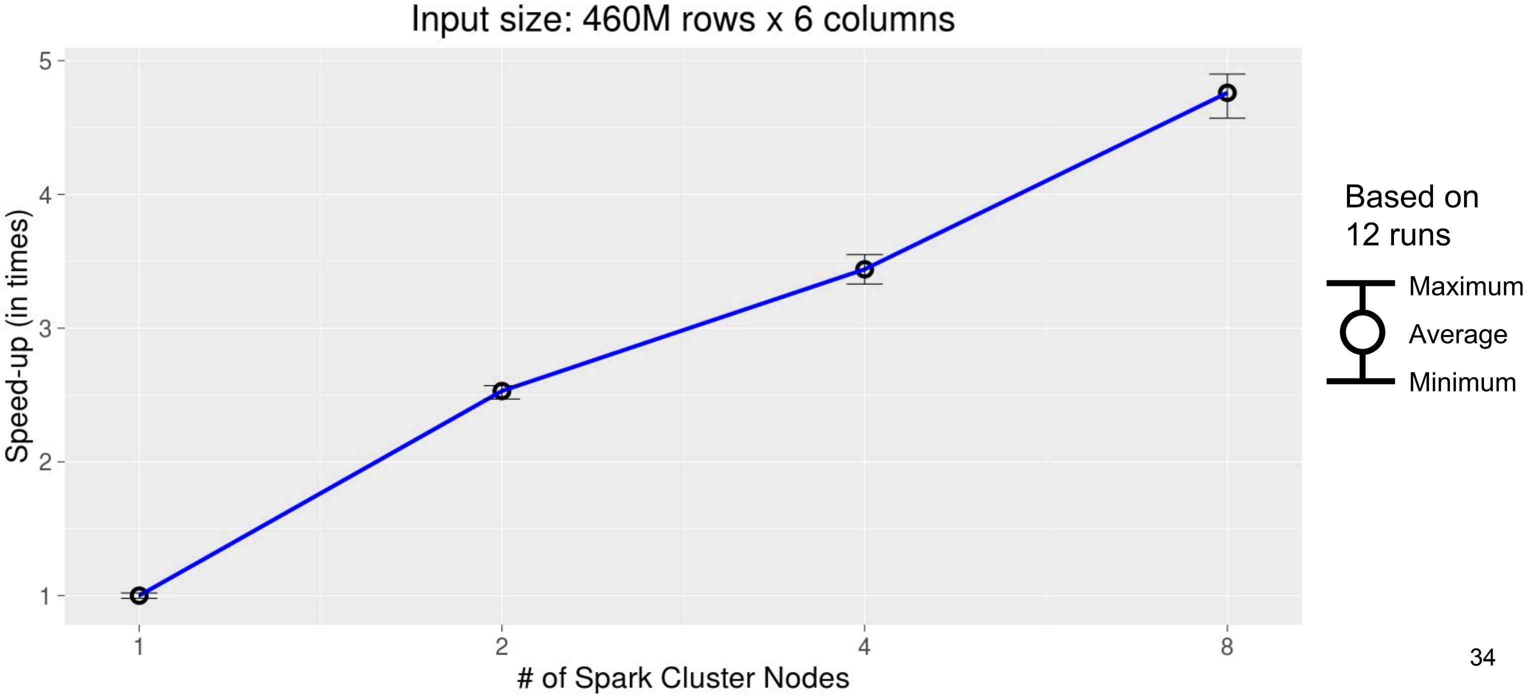
Experimental Setup (2)

- Workload: Bar graph (`ggplot(df, aes(x=Item))+geom_bar()`)



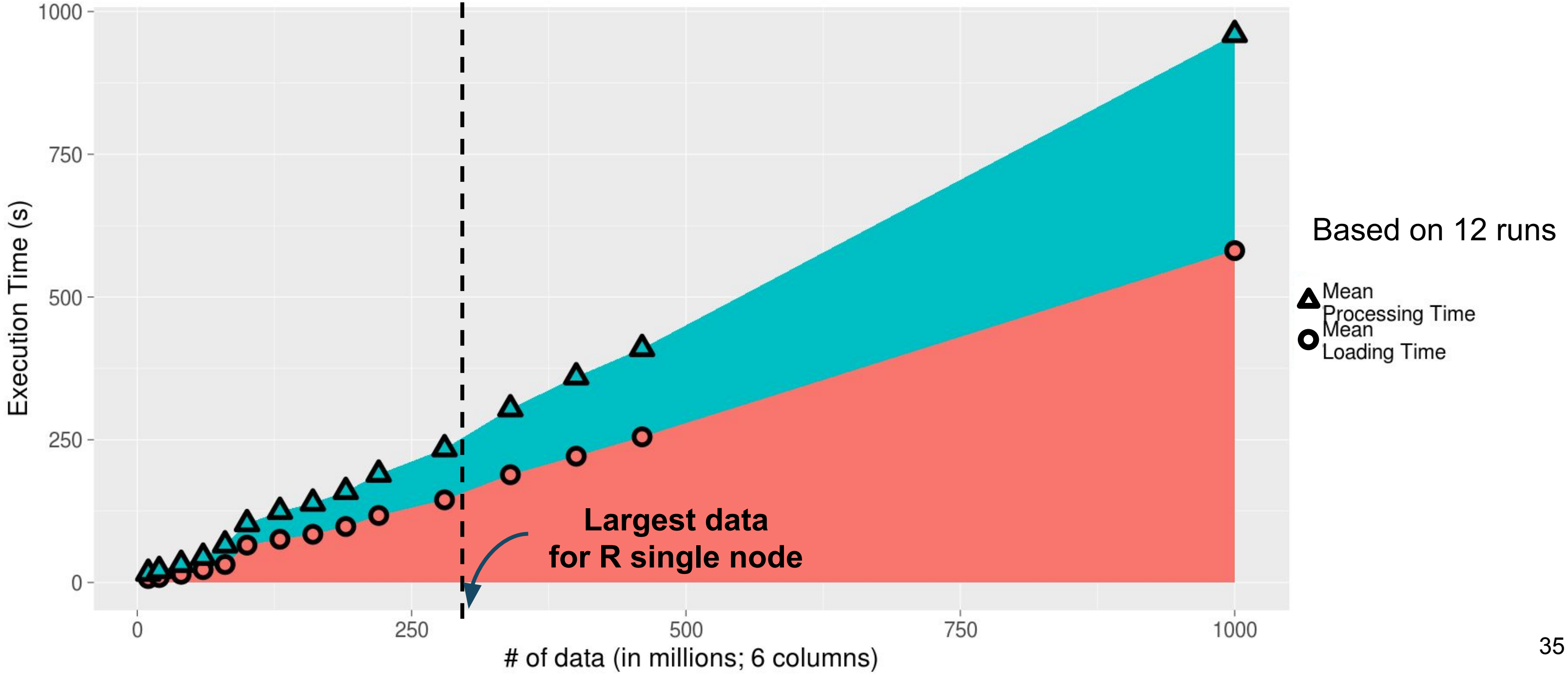
Performance: Scalability

Performance scales to the number of cluster nodes.



Performance: Varying Data Size

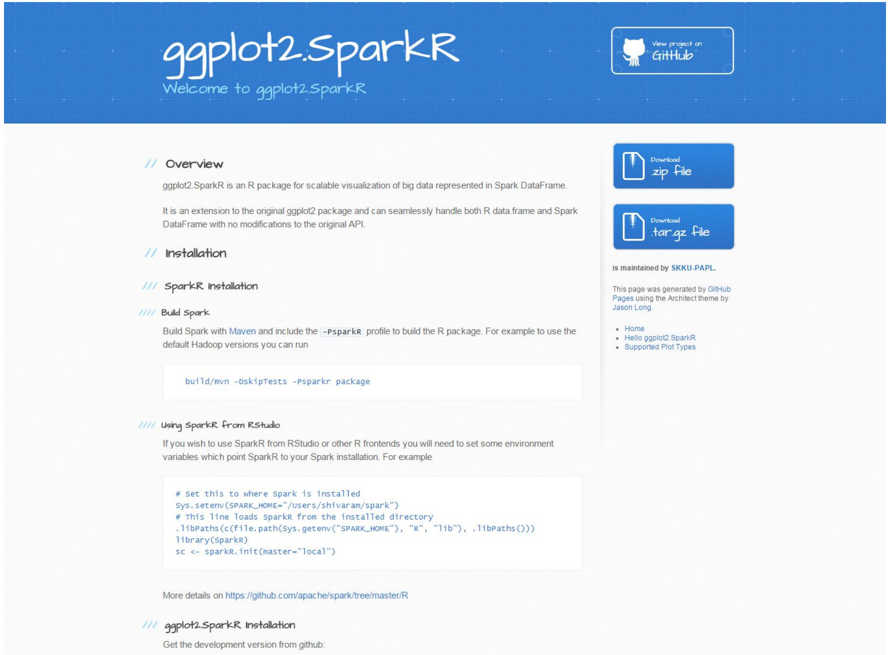
Throughput (inverse of slope) remains relatively stable.



Status & Plan

ggplot2.SparkR Project Page

Project page: http://skku-skt.github.io/ggplot2.SparkR



// Overview

ggplot2.SparkR is an R package for scalable visualization of big data represented in Spark DataFrame.

It is an extension to the original ggplot2 package and can seamlessly handle both R data.frame and Spark DataFrame with no modifications to the original API.

// Installation

/// SparkR Installation

//// Build Spark

Build Spark with Maven and include the `-PsparkR` profile to build the R package. For example to use the default Hadoop versions you can run

```
build/mvn -DskipTests -Psparkr package
```

//// Using SparkR from RStudio

If you wish to use SparkR from RStudio or other R frontends you will need to set some environment variables which point SparkR to your Spark installation. For example

```
# Set this to where Spark is installed
Sys.setenv(SPARK_HOME="/Users/shivaram/spark")
# This line loads SparkR from the installed directory
.libPaths(c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib"), .libPaths()))
library(SparkR)
sc <- sparkR.init(master="local")
```

More details on https://github.com/apache/spark/tree/master/R

/// ggplot2.SparkR Installation

Get the development version from github:

is maintained by SKKU-PAPL.

This page was generated by GitHub Pages using the Architect theme by Jason Long.

- Home
- Hello ggplot2.SparkR
- Supported Plot Types

To Report Bugs or Request Features

- Report using our github issue page

  https://github.com/SKKU-SKT/ggplot2.SparkR/issues

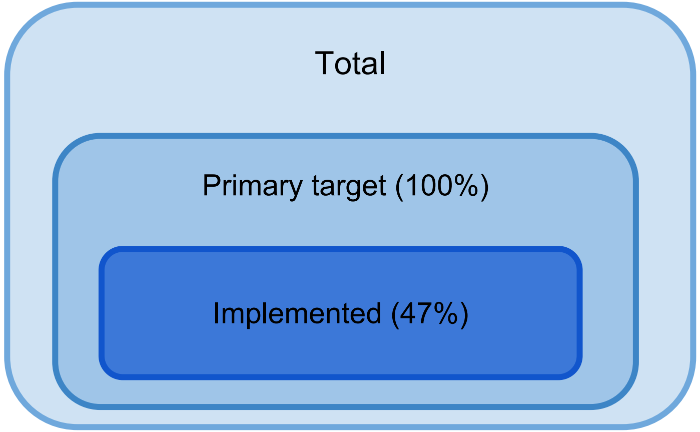
Or

- Email to the ggplot2.SparkR mailing list

  ggplot2-sparkr@googlegroups.com

API Coverage & Future Plan

- ggplot2 API Coverage
  - Total: 135
  - Primary target: 45* (100%)
  - Implemented: 21 (47%)



- Future Plan
  - Register the project to spark-packages.org (and CRAN)
  - Improve API coverage
  - Optimize performance

*Suitable for big data visualization

Summary: ggplot2.SparkR

- R package extending ggplot2 to take Spark DataFrame (as well as R data.frame) as input
- Scalable, easy to use, and readily deployable
- Feedback and contributions from Spark Community will be greatly appreciated.

THANK YOU.

https://github.com/SKKU-SKT/ggplot2.SparkR