SparkR

Big Data & R

DataFrames
Visualization
Libraries

+

Data

Background

Engine for large-scale data processing

Fast, Easy to Use

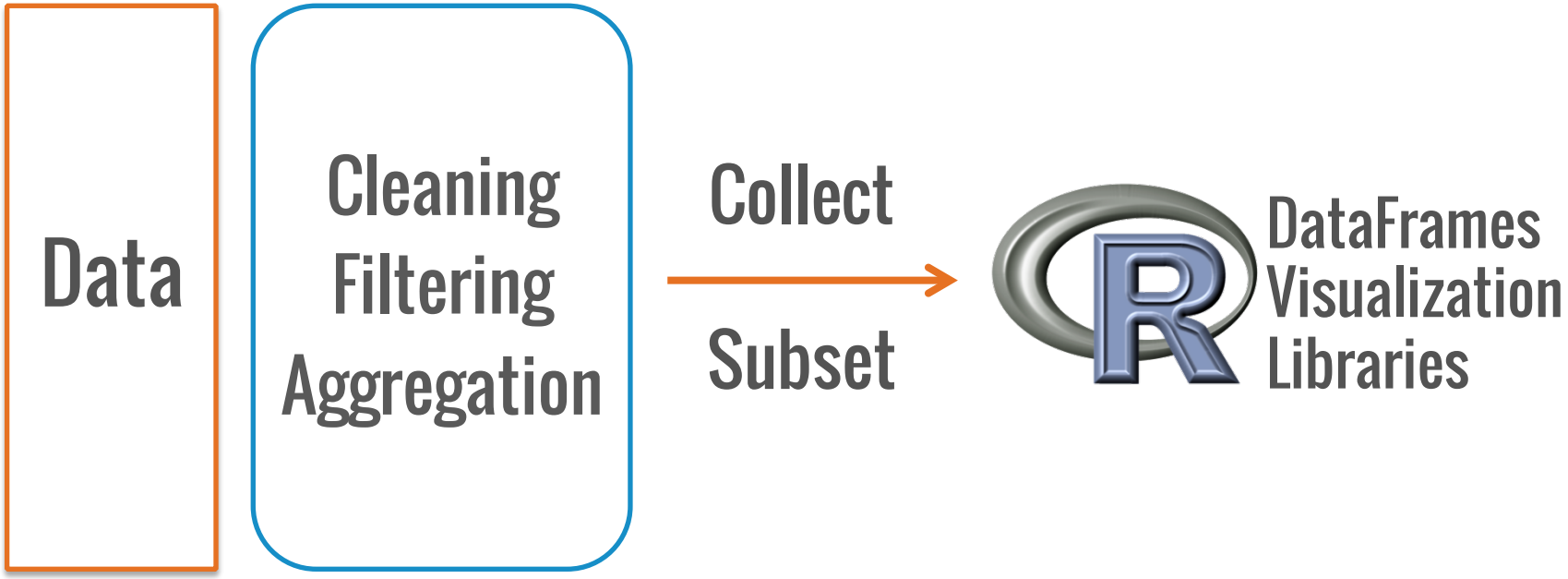
Runs Everywhere - EC2, YARN, Mesos

SparkR

Interactive Shell

Batch Scripts

Big Data Processing + R



SparkR DataFrames

High-level API for data manipulation

Read in CSV, JSON, JDBC etc.

dplyr-like syntax

Example


```
{"name":"Michael", "age":29}
{"name":"Andy", "age":30}
{"name":"Justin", "age":19}
{"name":"Bob", "age":22}
{"name":"Chris", "age":28}
{"name":"Garth", "age":36}
{"name":"Tasha", "age":24}
{"name":"Mac", "age":30}
{"name":"Neil", "age":32}
```

Example

```
people <- read.df(
    “hdfs://people.json”,
    “json”)

avgAge <- select(
    df,
    avg(df$age))

collect(avgAge)
```

Read input from HDFS

Collect to data.frame

DataFrame API

Filtering Data

- select, `$`, where, filter

Aggregating Data

- groupBy, summarize, arrange

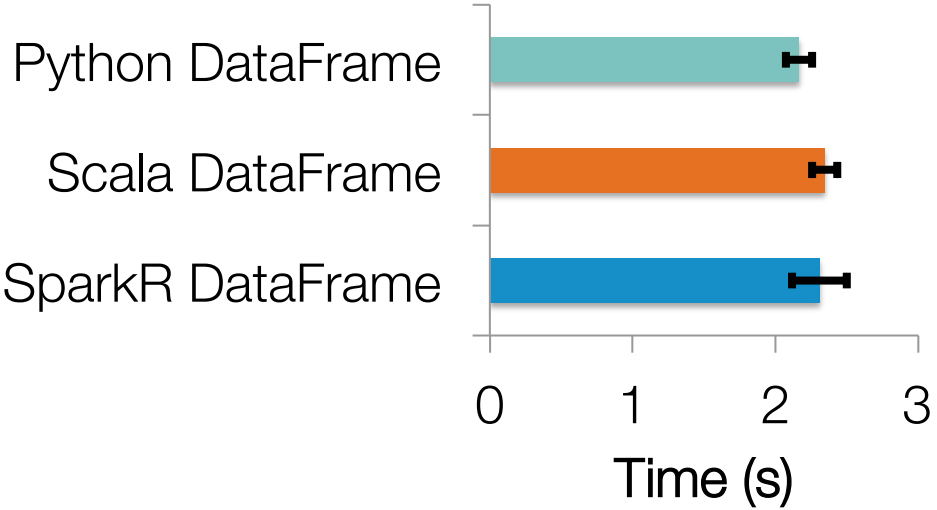
Input/Output

- read.df, write.df, sql

SparkR DataFrames

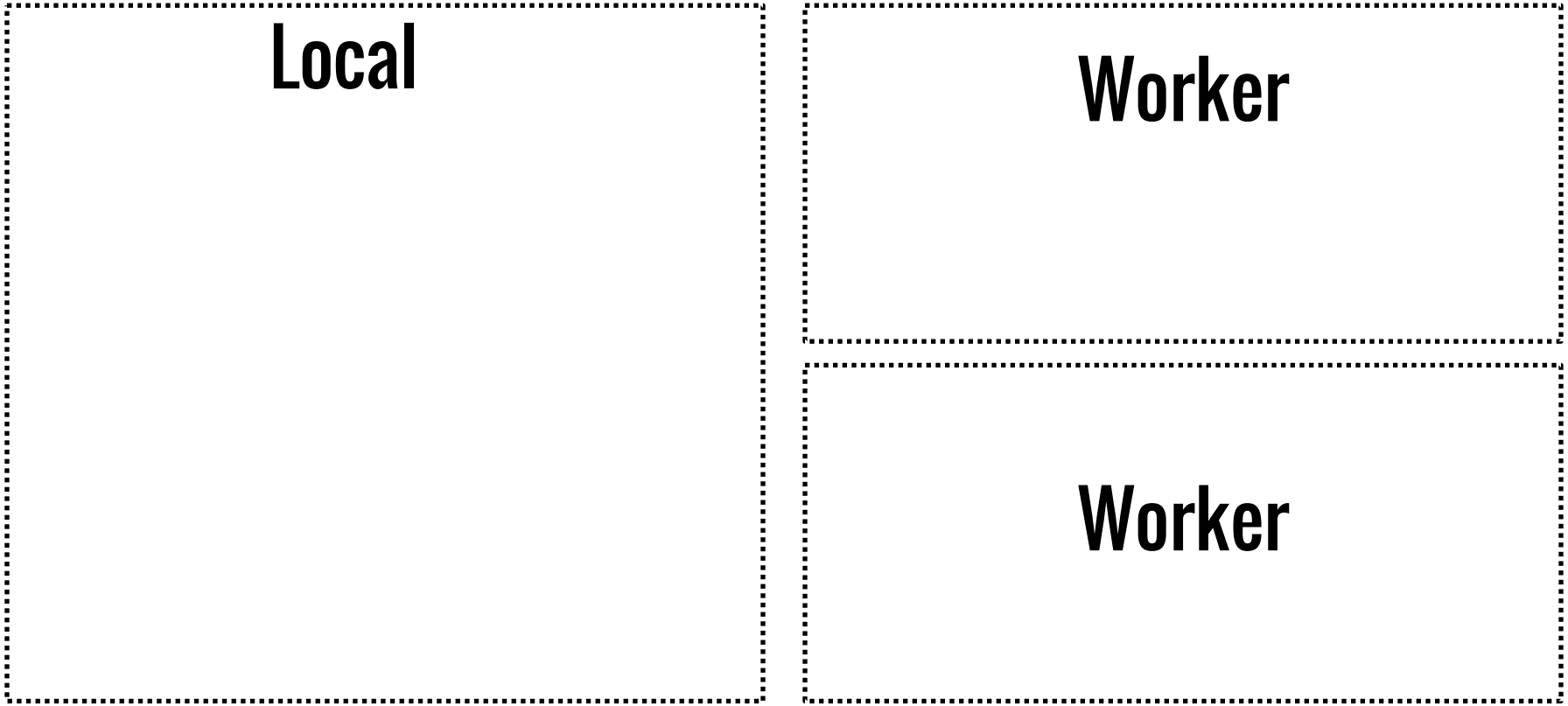
Query Planning

SQL Optimizations

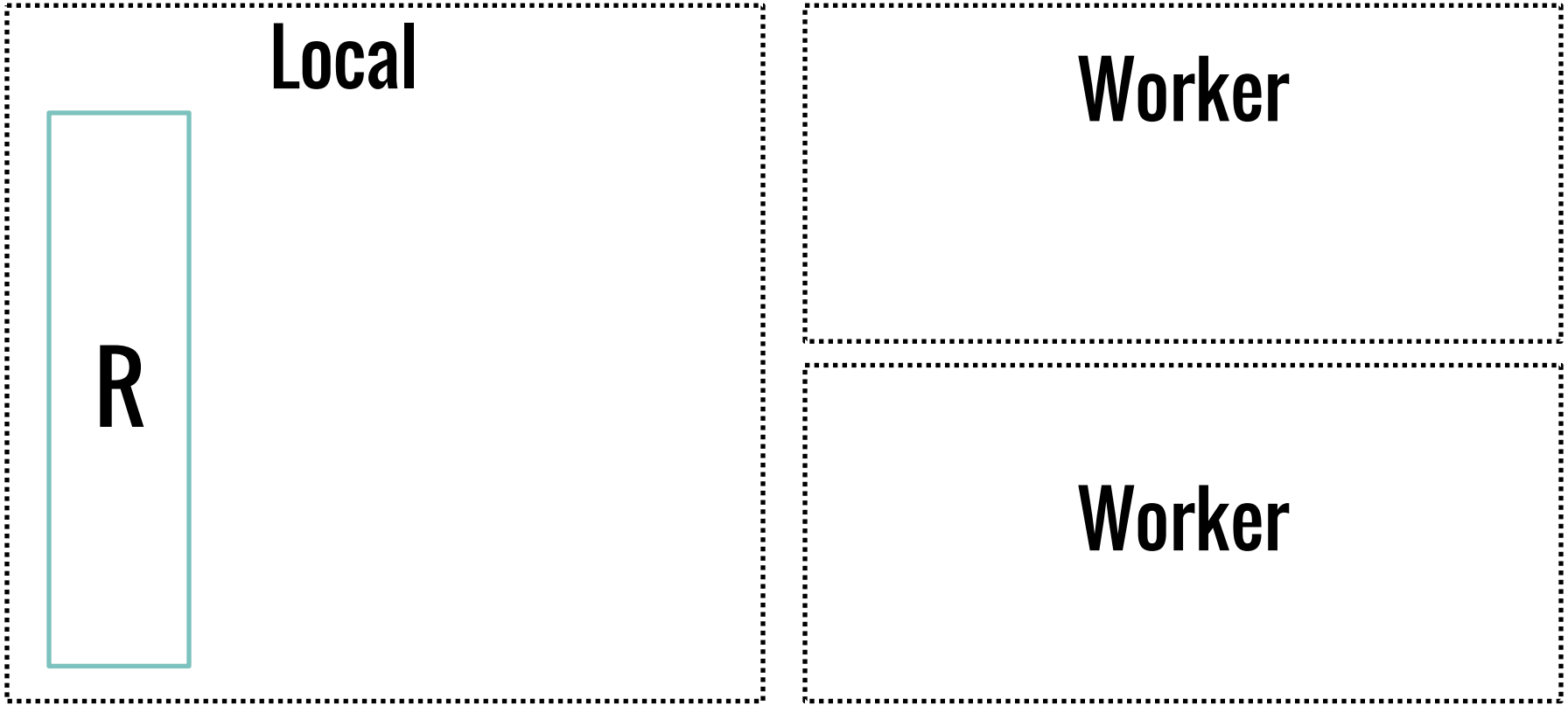


Architecture

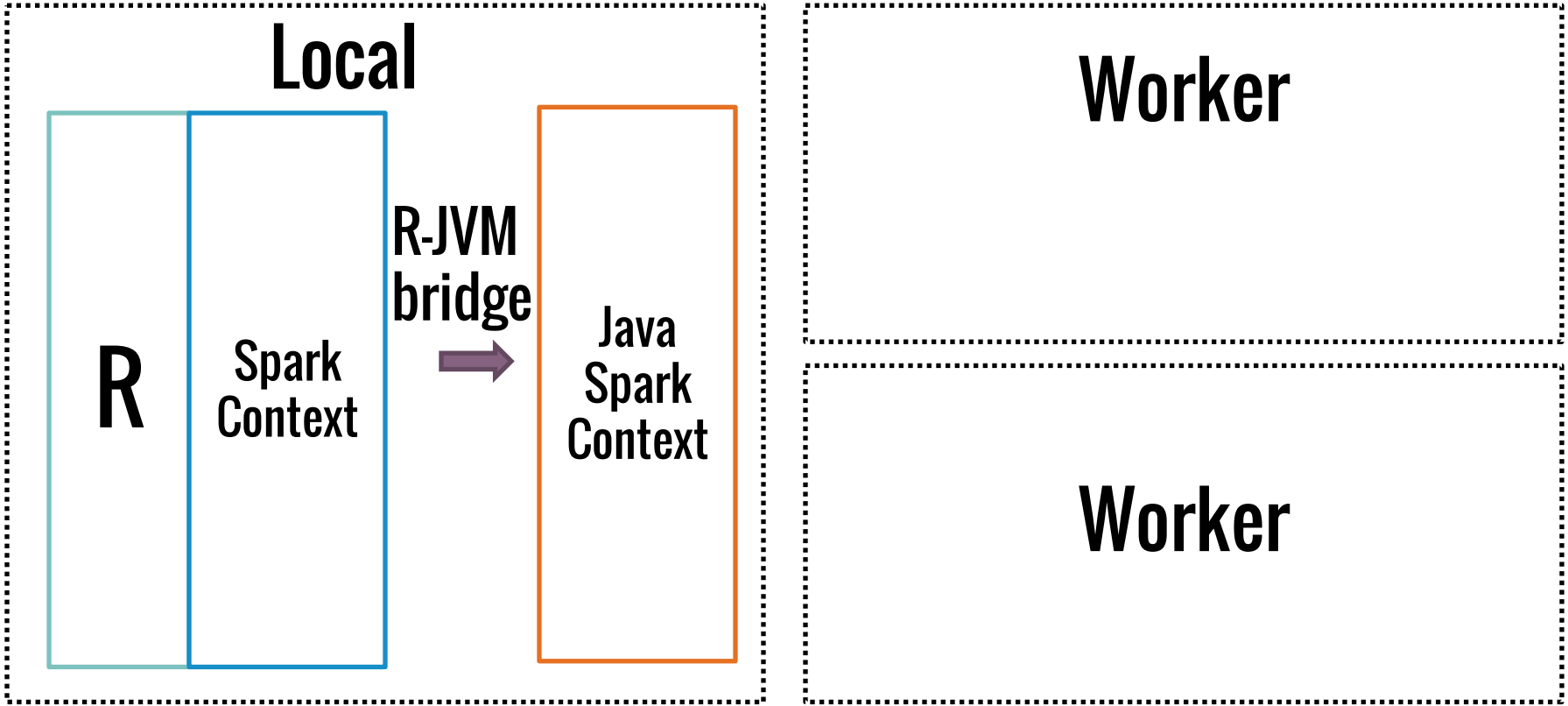
Architecture



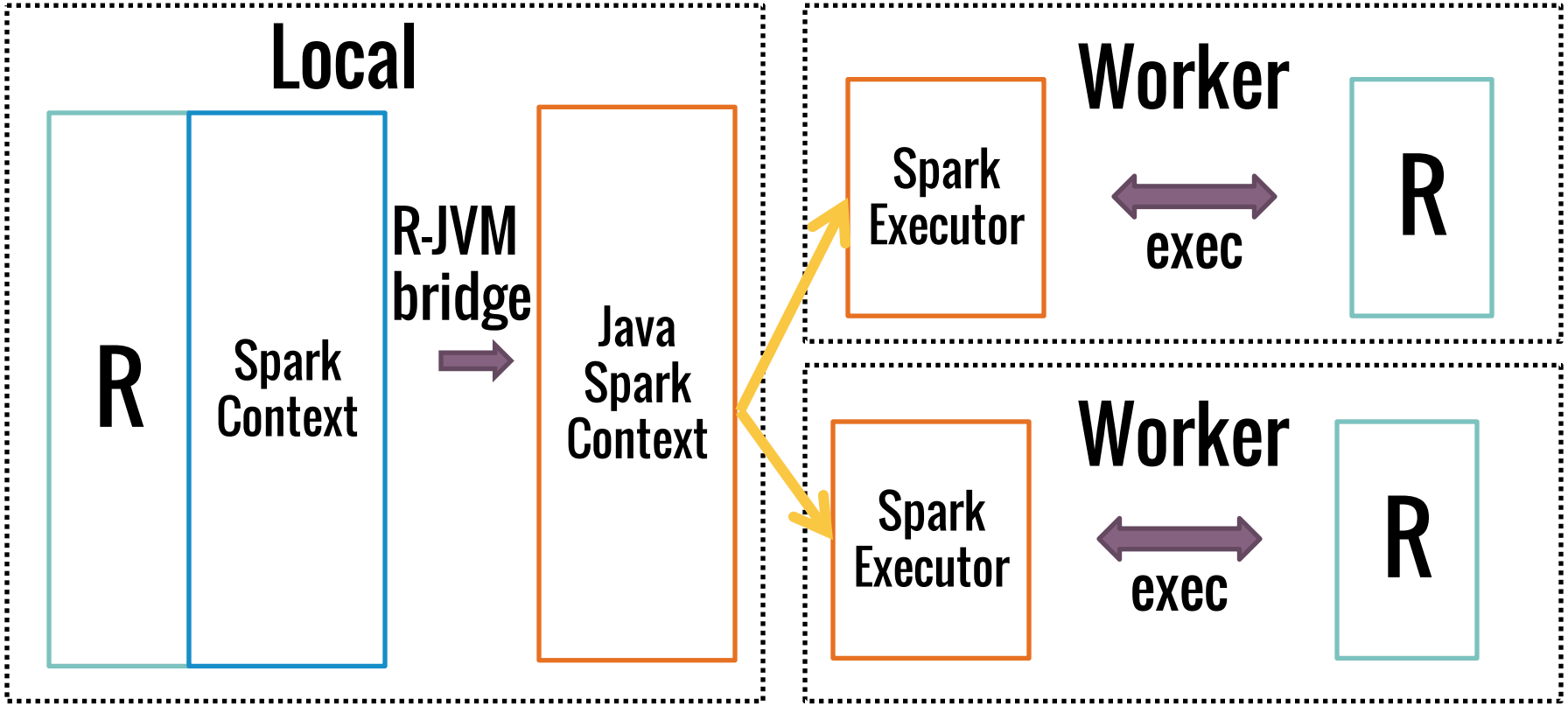
Architecture



Architecture



Architecture



Running SparkR Locally

Download from http://spark.apache.org/ (>1.4.0)

./bin/sparkR or RStudio

Useful for learning SparkR, demonstrations

Big Data & R

Big Data
Small Learning

Partition
Aggregate

Large Scale
Machine Learning

SparkR:
Unified approach

SparkR

Big data processing from R

DataFrames in Spark 1.4

Future: Large Scale ML & more