

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190

SparkR (R on Spark)

- [Overview](#)
- [SparkDataFrame](#)
 - [Starting Up: SparkSession](#)
 - [Starting Up from RStudio](#)
 - [Creating SparkDataFrames](#)
 - [From local data frames](#)
 - [From Data Sources](#)
 - [From Hive tables](#)
 - [SparkDataFrame Operations](#)
 - [Selecting rows, columns](#)
 - [Grouping, Aggregation](#)
 - [Operating on Columns](#)
 - [Applying User-Defined Function](#)
 - [Run a given function on a large dataset using dapply or dapplyCollect](#)
 - [dapply](#)
 - [dapplyCollect](#)
 - [Run a given function on a large dataset grouping by input column\(s\) and using gapply or gapplyCollect](#)
 - [gapply](#)
 - [gapplyCollect](#)
 - [Run local R functions distributed using spark.lapply](#)
 - [spark.lapply](#)
 - [Running SQL Queries from SparkR](#)
- [Machine Learning](#)
 - [Algorithms](#)
 - [Classification](#)
 - [Regression](#)
 - [Tree](#)
 - [Clustering](#)
 - [Collaborative Filtering](#)
 - [Frequent Pattern Mining](#)
 - [Statistics](#)
 - [Model persistence](#)
- [Data type mapping between R and Spark](#)
- [Structured Streaming](#)
- [R Function Name Conflicts](#)
- [Migration Guide](#)