SparkR in one slide

DataFrame is the main data structure
    Translated to RDD[Row] behind the scenes. Columns have names and data type.

All local R functions and libraries available (locally on the driver only for now).

Connectors to data sources (csv, json, jdbc, parquet, hive, mysql, postgresql, hdfs, aws s3, h2, …).

partitions, lazy eval, query planning, predicate pushdown, fault-tolerance, in-memory caching, compression, sql.

Stage of maturity: use for filtering, selecting, and aggregating your BIG data down to small or medium data which you can work with locally.

Some machine learning algorithms and stats are available on the BIG data, more will be added as the SparkR project grows.

Can work in REPL/shell or Notebook (eg. RStudio, Jupyter, DBCE).