SPARK+AI SUMMIT 2020

Organized by databricks

How R Developers Can Build and Share Data and AI Applications that Scale with Databricks and Rstudio Connect

James Blair, Solutions Engineer, RStudio PBC

Rafi Kurlansik, Sr. Solutions Architect, Databricks

Agenda

Rafi Kurlansik, Databricks

Building Scalable R and Shiny apps with RStudio and Databricks

James Blair, RStudio PBC

Deploying Scalable Shiny apps with RStudio Connect and Databricks

Benchmarking performance of Shiny connections to Spark

How to scale R and Shiny with RStudio and Databricks

SPARK+AI SUMMIT

How can we open up the data lake to R users?

Imagine trying to do so with traditional R development...

- ## Typical development patterns
  - Local
  - Cloud / On Prem VM
- ## Challenges with big data
  - Server memory - can only process so much data in the app itself before crashing R
  - Performance - even on a powerful VM, eventually see our app get less responsive as we reach 100+ GBs
  - Managing big data infrastructure - app value must be higher to justify the energy investment

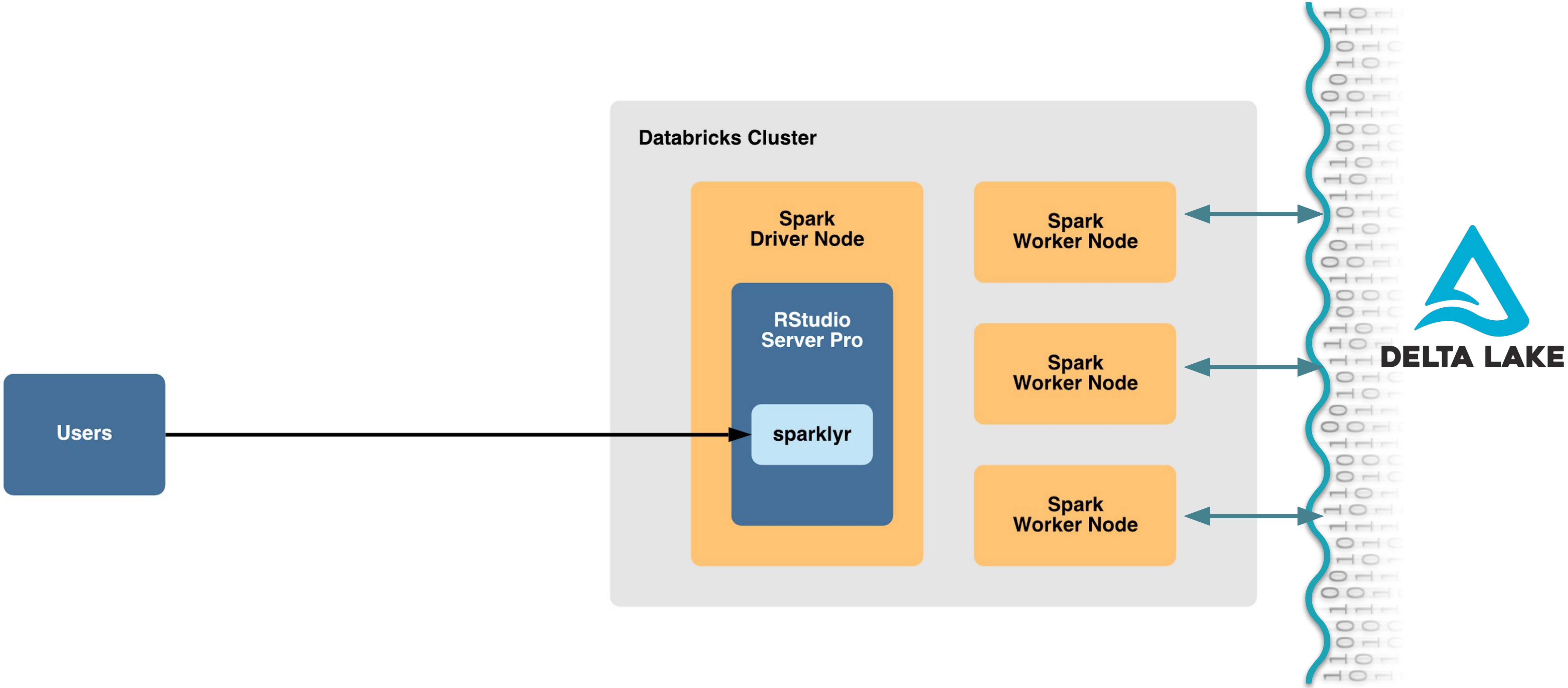
If only there was a technology with a familiar API in R that let our app scale to process 100s of GBs...

Scale R Apps with Databricks and RStudio

Databricks Spark, RStudio IDE

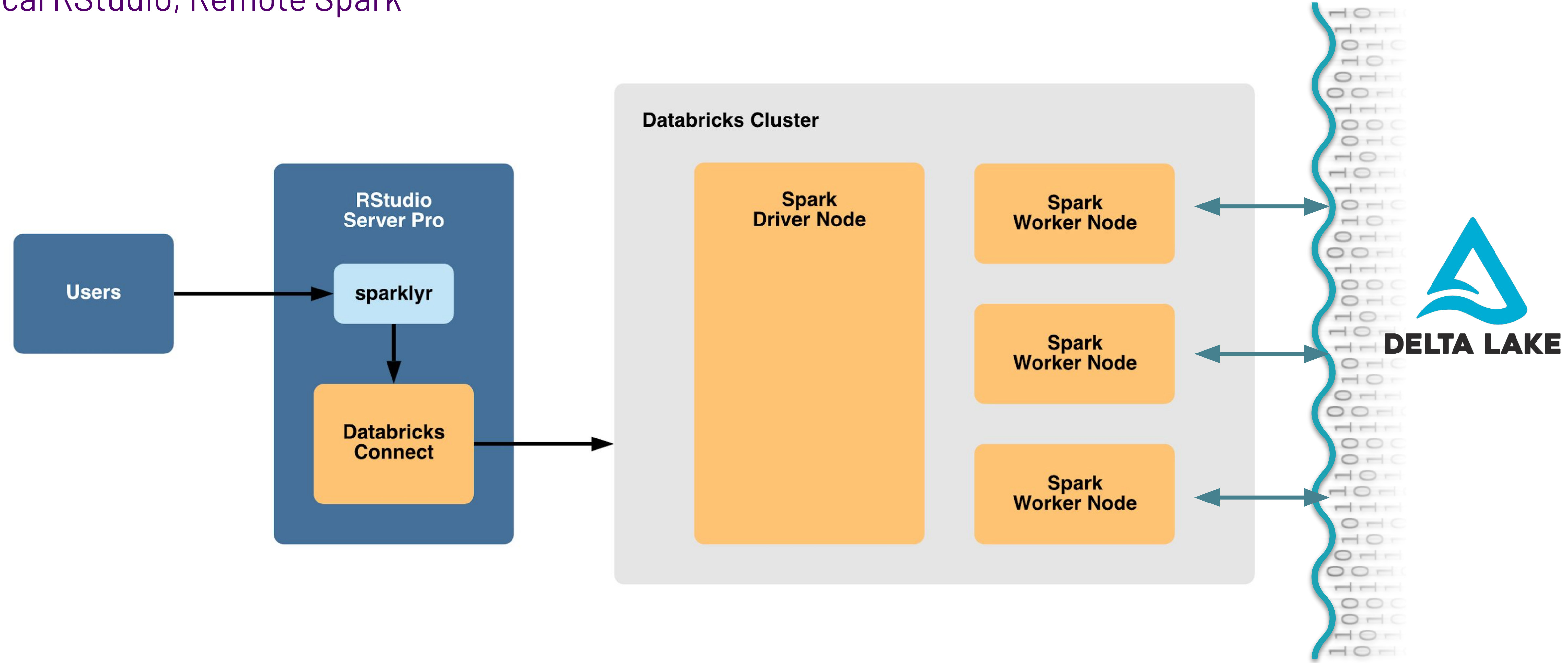
- Development Patterns
  - Hosted RStudio Server (Pro) on Databricks Cluster
  - RStudio with remote Spark access using Databricks Connect
- Overcoming challenges with big data
  - Auto-scaling Databricks Spark Clusters - dynamically respond to accommodate larger data processing tasks
  - Consistently fast performance with Delta Lake and Databricks Runtime
  - Managed service allows data teams to focus on building data products, not maintaining infrastructure

Hosted RStudio Server Pro on Databricks



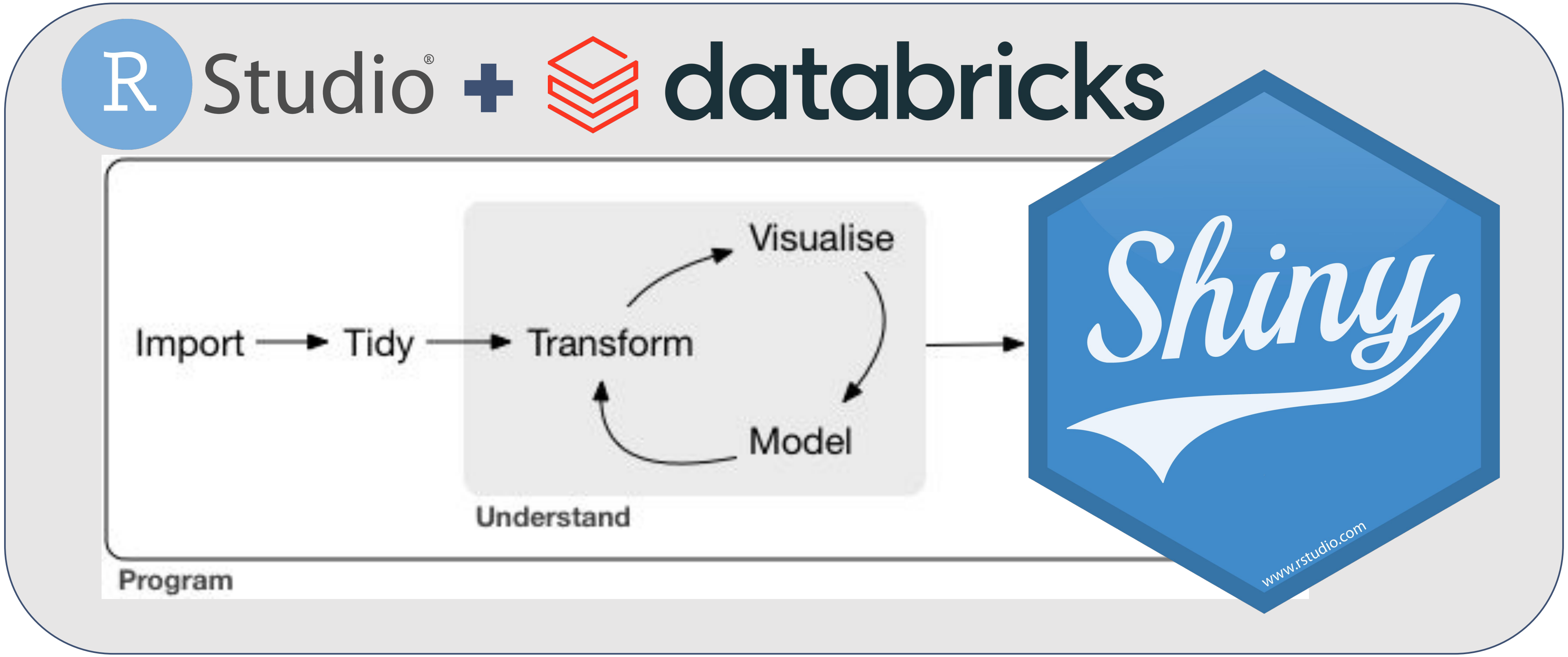
RStudio with Databricks Connect

Local RStudio, Remote Spark



Sharing Scalable Shiny Apps

The Data Science Process

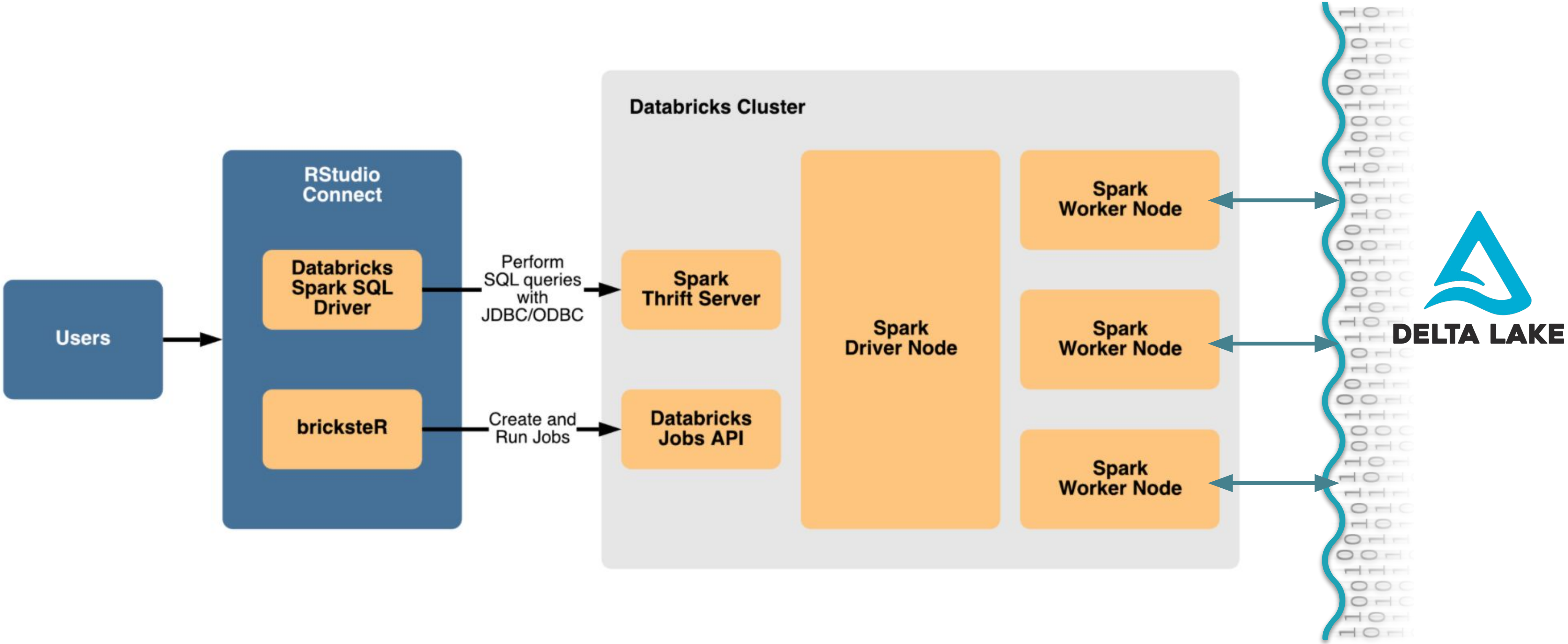


Shiny and Spark: A cautionary tale

ODBC to the Rescue

- The R + ODBC toolchain is robust and stable
- As performant as a native Spark connection
- Easy to migrate code from sparklyr to ODBC
- Spark still does all of the computation
- Databricks provides an optimized Spark ODBC driver

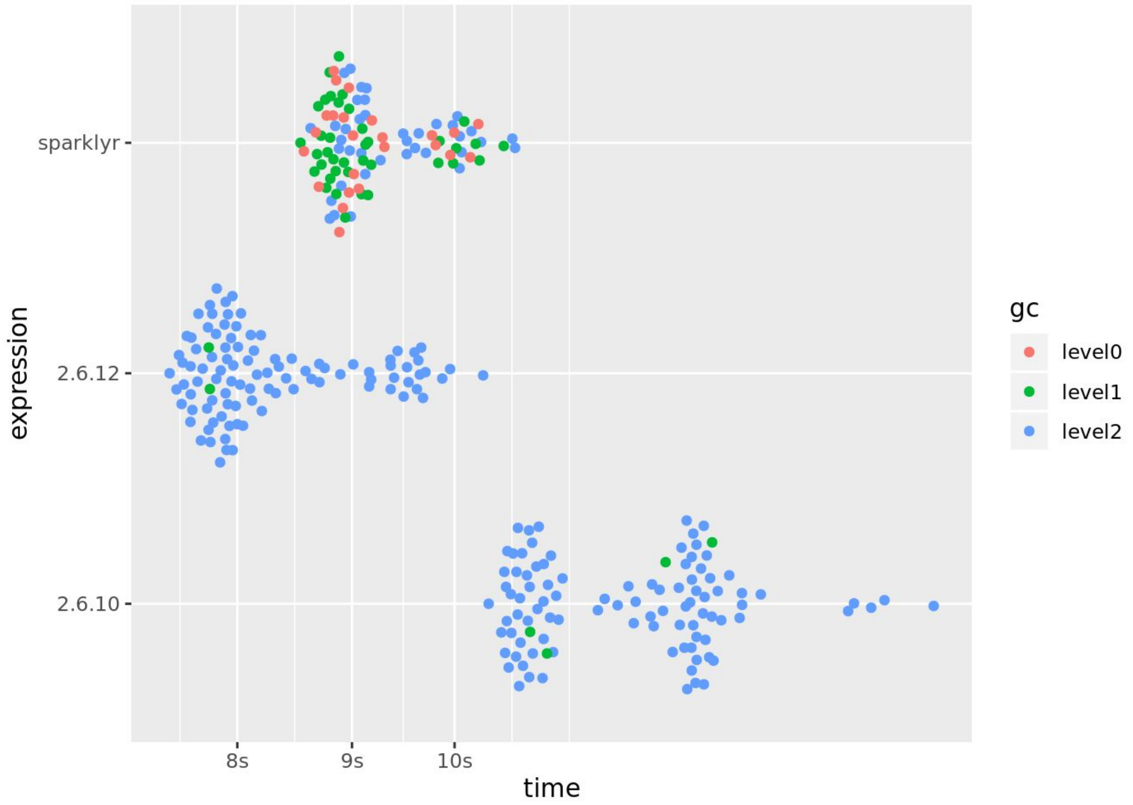
ODBC with RStudio Connect



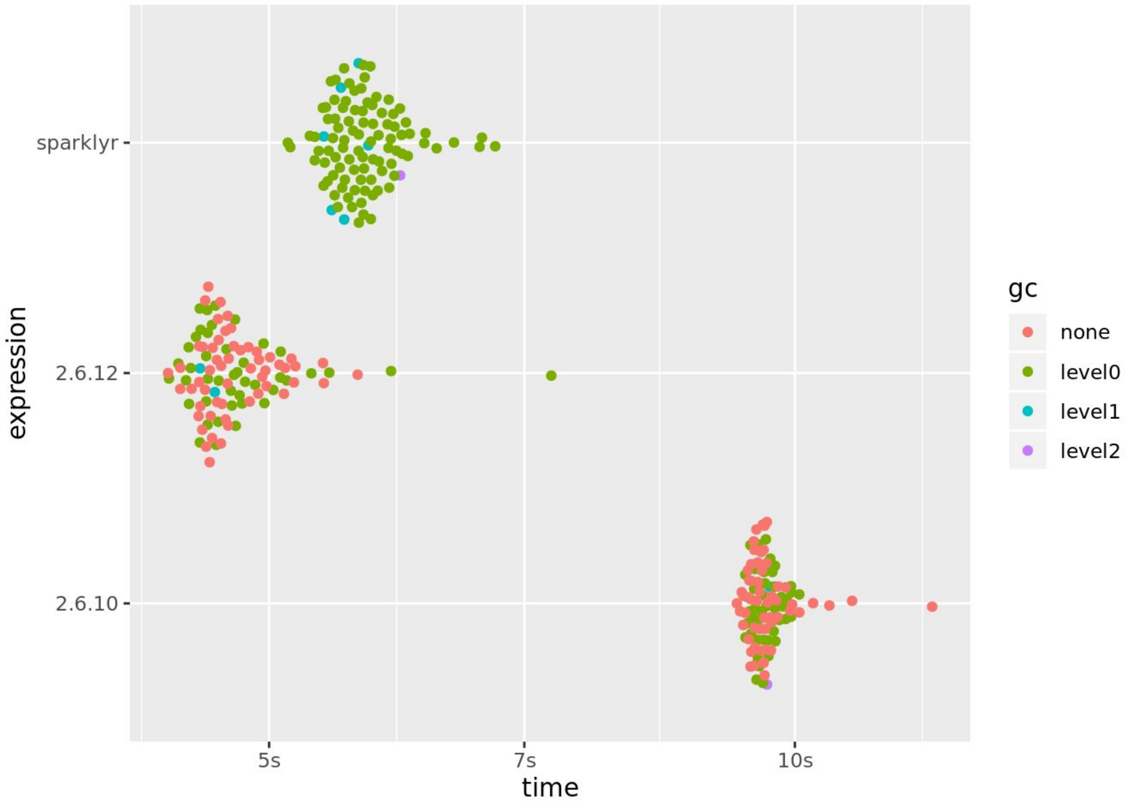
ODBC Performance

Comparing sparklyr against two versions of the Databricks ODBC/JDBC Driver

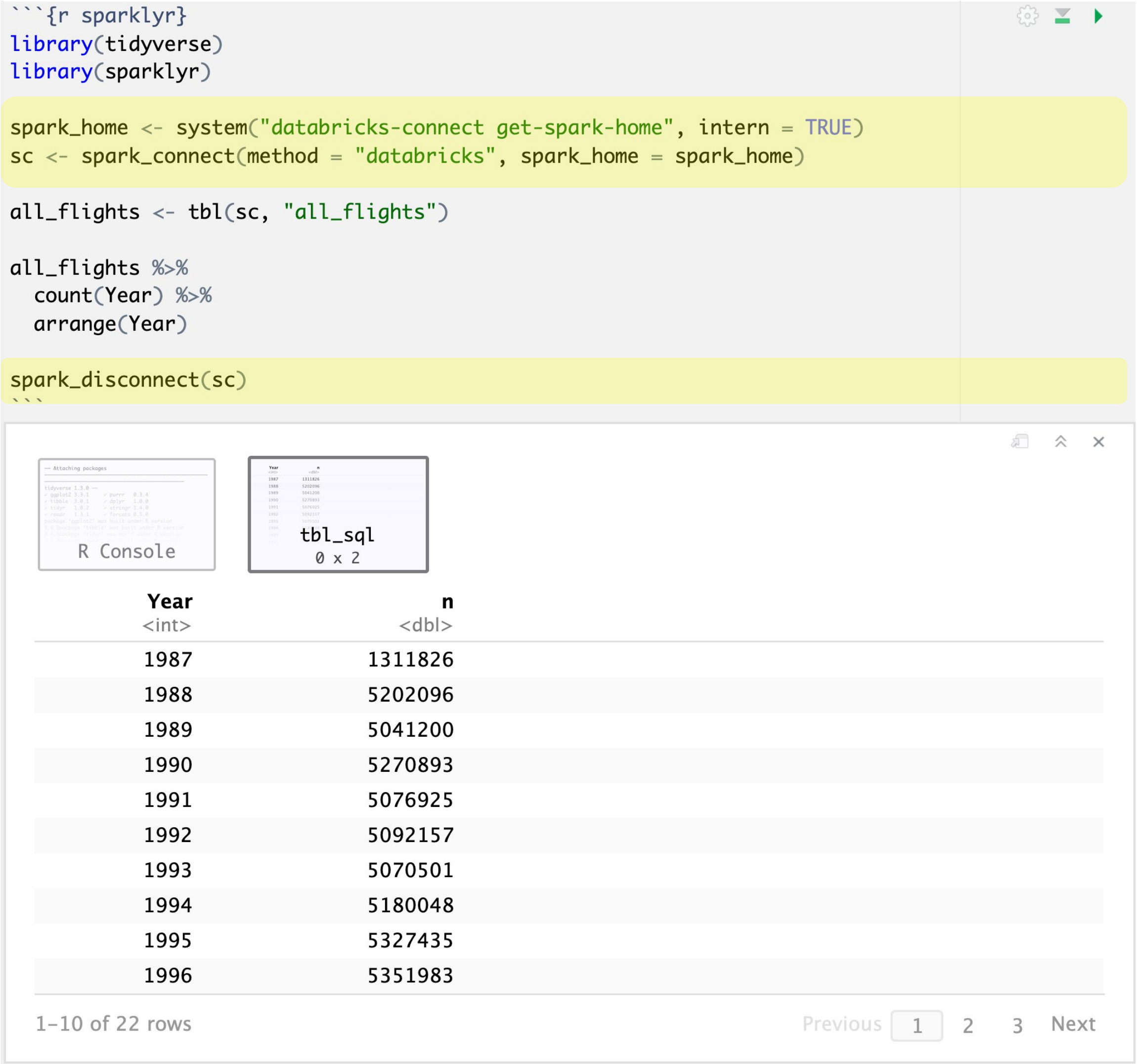
Collecting



Joins



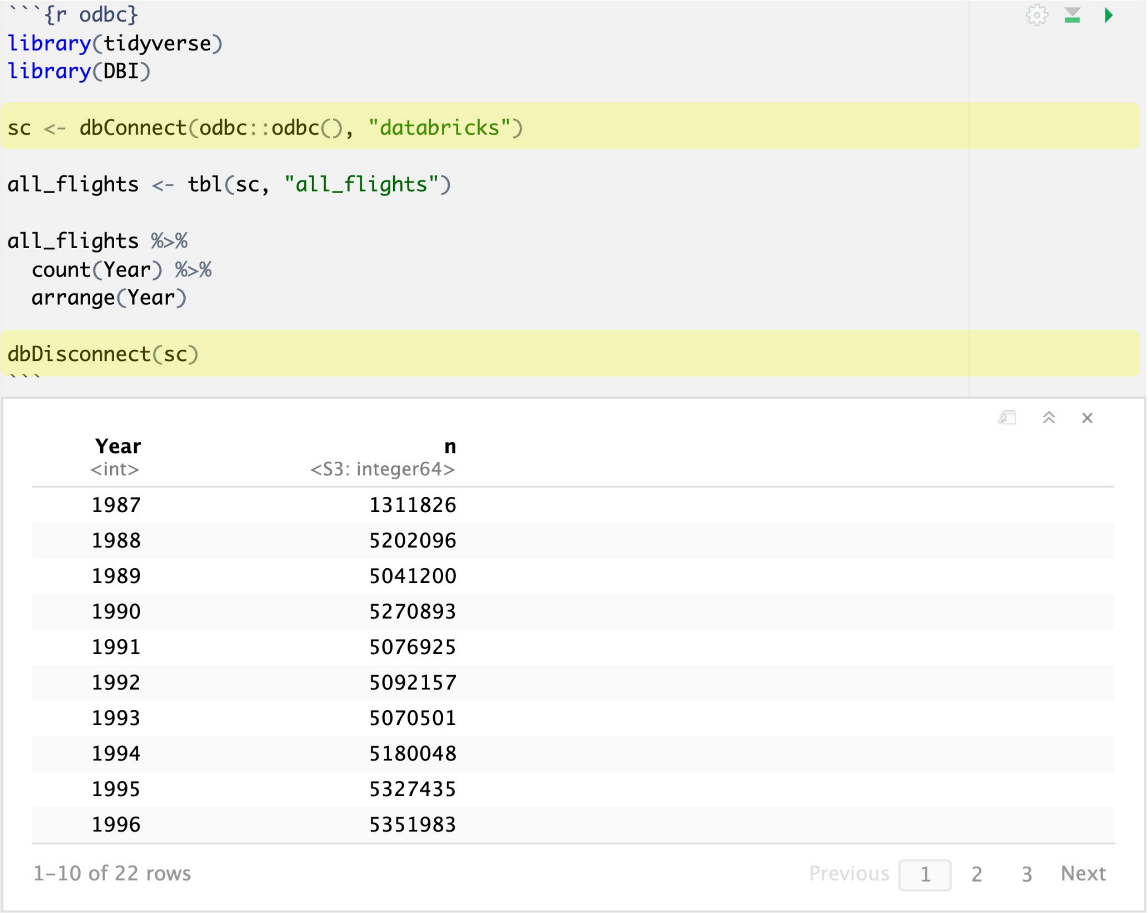
Sparklyr to ODBC



~~~
```{r sparklyr}
library(tidyverse)
library(sparklyr)

spark_home <- system("databricks-connect get-spark-home", intern = TRUE)
sc <- spark_connect(method = "databricks", spark_home = spark_home)

all_flights <- tbl(sc, "all_flights")

all_flights %>%
  count(Year) %>%
  arrange(Year)

spark_disconnect(sc)
```
~~~

Year <int>	n <dbl>
1987	1311826
1988	5202096
1989	5041200
1990	5270893
1991	5076925
1992	5092157
1993	5070501
1994	5180048
1995	5327435
1996	5351983

1-10 of 22 rows



~~~
```{r odbc}
library(tidyverse)
library(DBI)

sc <- dbConnect(odbc::odbc(), "databricks")

all_flights <- tbl(sc, "all_flights")

all_flights %>%
  count(Year) %>%
  arrange(Year)

dbDisconnect(sc)
```
~~~

Year <int>	n <S3: integer64>
1987	1311826
1988	5202096
1989	5041200
1990	5270893
1991	5076925
1992	5092157
1993	5070501
1994	5180048
1995	5327435
1996	5351983

1-10 of 22 rows

Conclusion

RStudio + databricks

Develop at scale

- Interactive data analysis with SparkSQL
    - sparklyr
    - ODBC
- Other Spark APIs
    - sparklyr

Deploy at scale

- Interactive data analysis with SparkSQL
    - Shiny with ODBC
- Other Spark APIs
    - ¯_(ツ)_/¯
    - Deploy models with MLflow?
    - Submit individual commands with Databricks REST API 1.2?
    - Run sparklyr jobs from RStudio on Databricks with bricksteR?
    - Stay tuned....

Additional Resources

Documentation

- Hosted RStudio on Databricks
- Databricks Connect
- ODBC
- ODBC Configuration
- RStudio Connect
- Sparklyr

Related Repos

- blairj09-talks/spark-summit-2020
- RafiKurlansik/bricksteR
- delta-io/delta
- sparklyr/sparklyr

Feedback

Your feedback is important to us.

Don’t forget to rate and review the sessions.

SPARK+AI SUMMIT 2020

Organized by databricks