Processing Large Data in R Using Apache Spark

Hossein Falaki

June 2017

databricks

About me

- Software Engineer at Databricks Inc.
- Data Scientist at Apple Siri
- Started using Spark since 0.6
- Developed first version of Apache Spark CSV data source
- Developed Databricks R Notebooks
- Currently focusing on R experience at Databricks

About Databricks

TEAM

Creators of Spark (now Apache Spark) at UC Berkeley in 2009

MISSION

Making big data simple

PRODUCT

Unified analytics platform

Outline

- Our view of R in enterprise
- Databricks data pipeline
- How Databricks enables R usage in enterprise
- How we use Databricks to do data science with R at Databricks
- Other use cases

Today: R usage in enterprise

- R is popular among advanced users (scientists & statisticians)
  - Sometimes hundreds of R users in one organization
- However, R is rarely productionized
  - R scripts are not executed against most of the data
  - In many cases R users are in disconnected pockets
  - BI tools and power point slides are used for broad consumption
  - Algorithms are re-implemented by software/data engineers for production

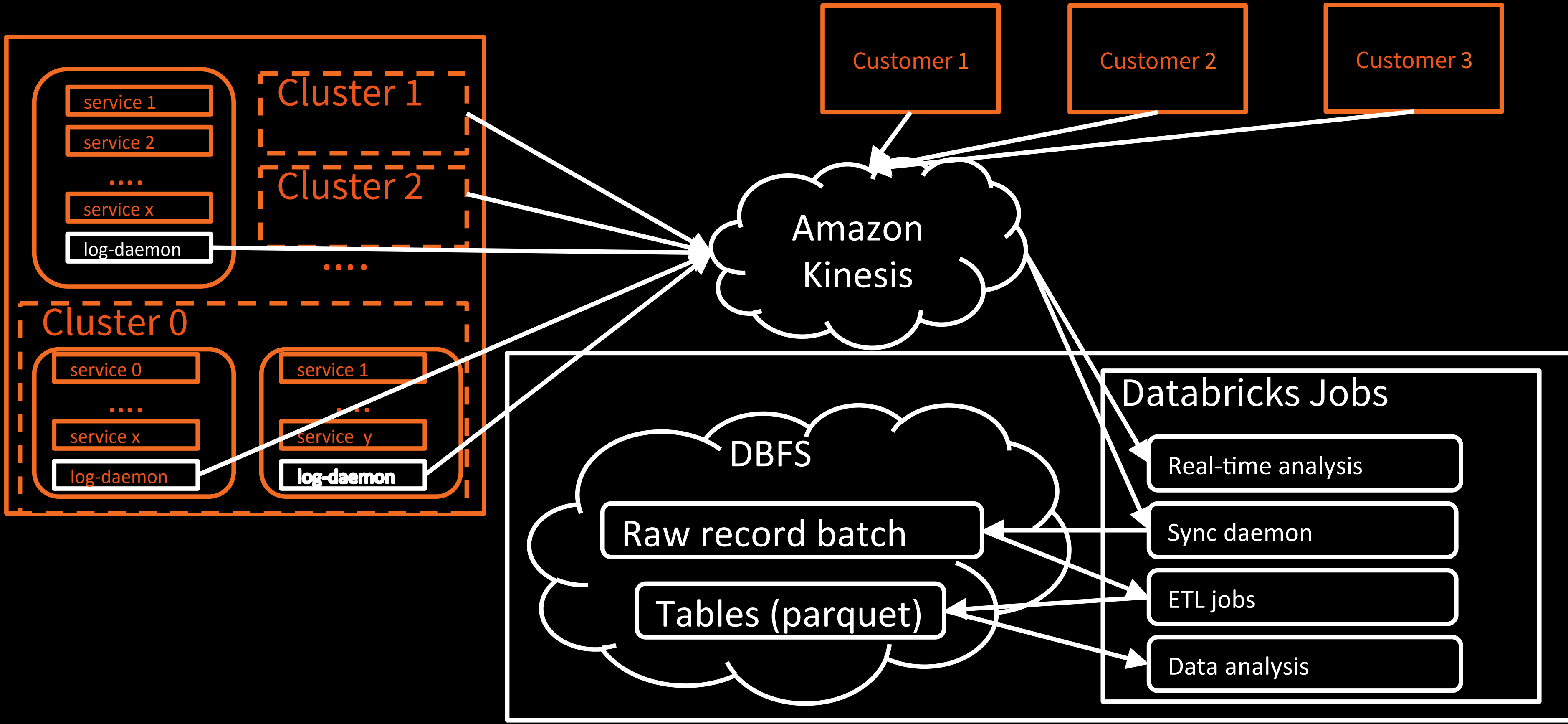
Ideal: R usage in enterprise

- Expose R to more individuals and teams
  - Consume
  - Run
  - Develop
- Expose more data to R code
  - R users can run their code on all of data: no sampling or pre-aggregation
  - R code is executed constantly as jobs

How to get from current to ideal

- Scalability
- Data access
- Collaboration
- Reproducibility
- Sharing and publishing
- Deploying models built in R to production
- Existing enterprise requirements

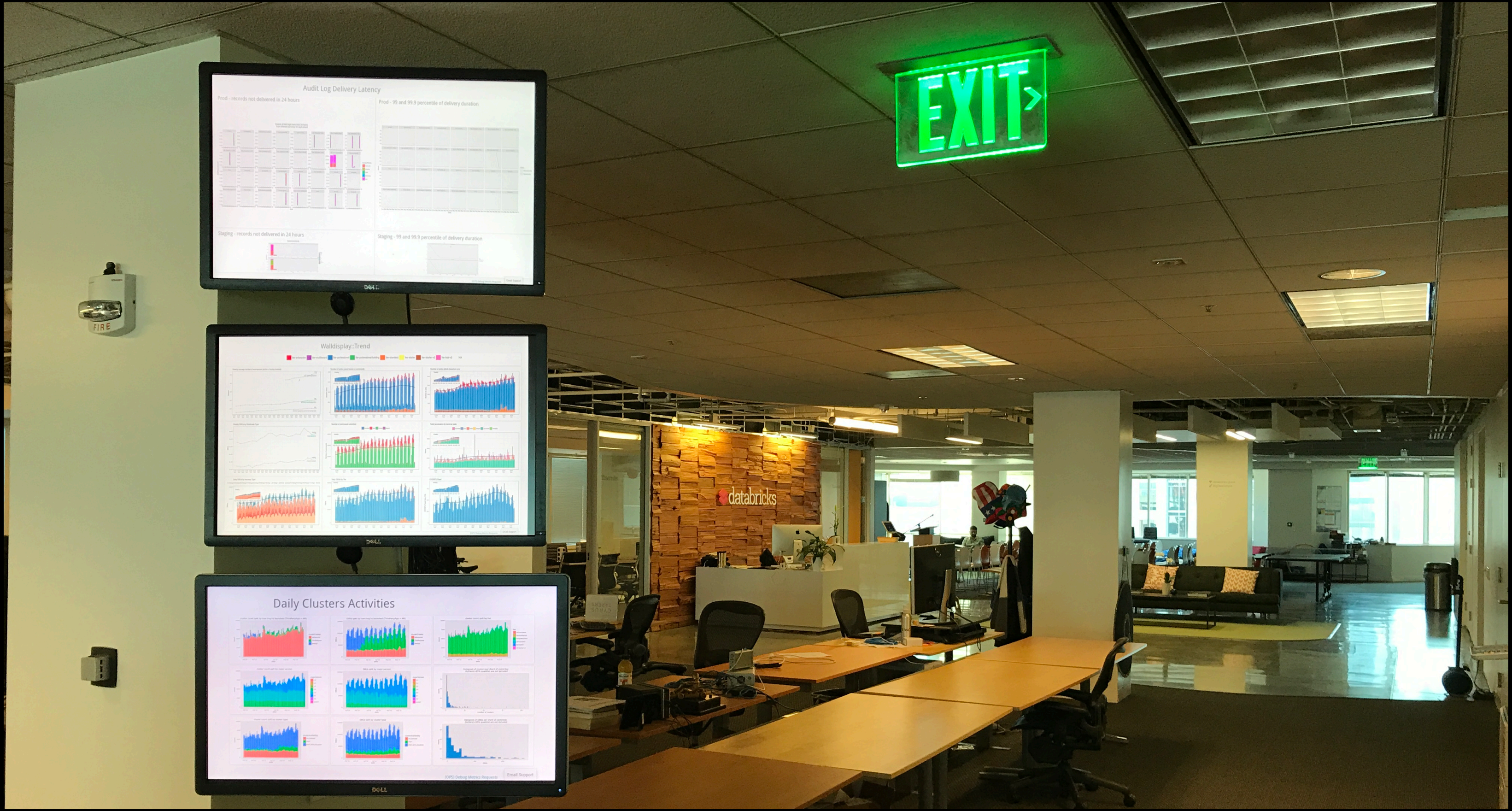
Example: Databricks data pipeline



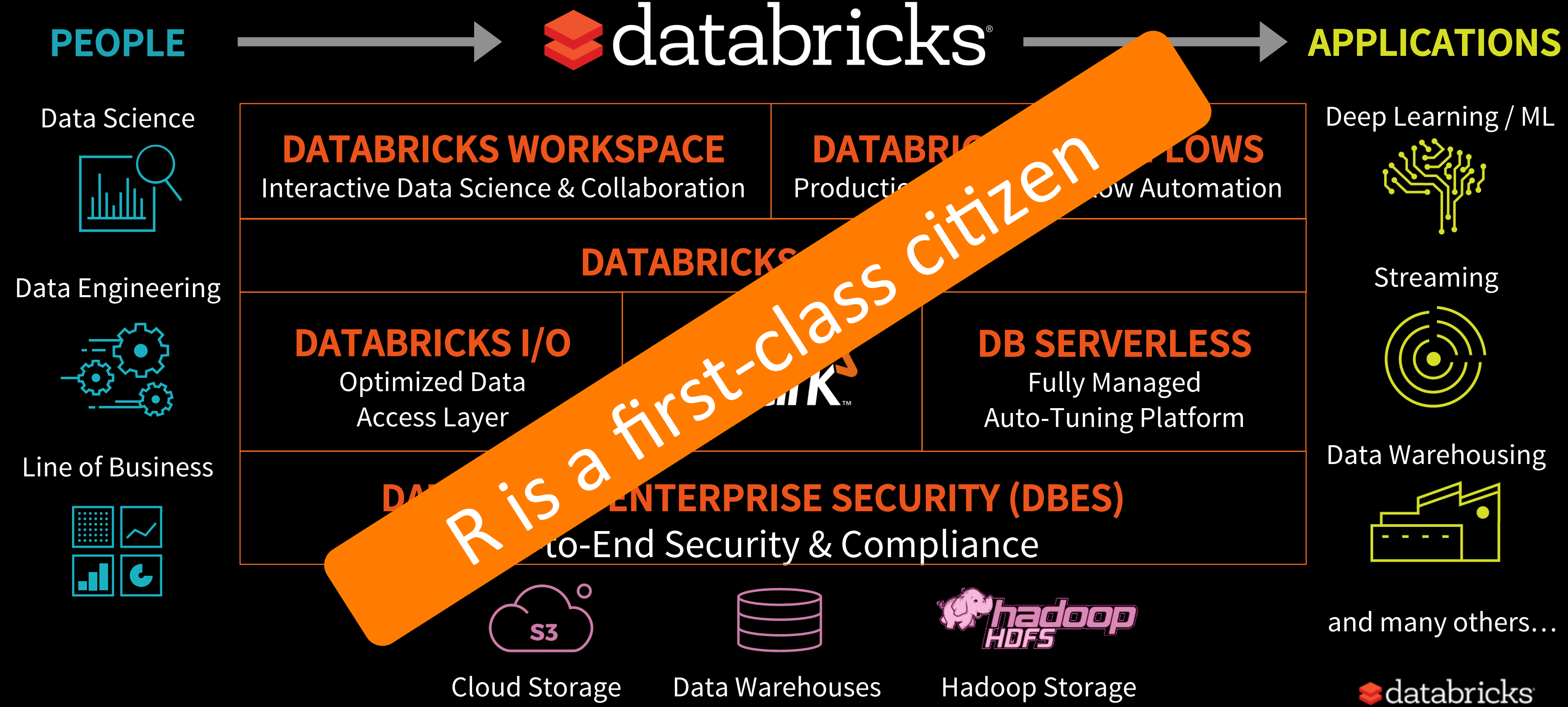
We heavily use R at Databricks

- Data scientists, some engineers and PM use R as primary language to analyze usage logs
- Daily, weekly and monthly reports are generated using R
- Production dashboards on the walls built in R
- Interactive dashboards for executive team
- Deep-dive investigations and reports are built with R notebooks
- Machine learning for sales and marketing lead scoring is mostly done in R

R used at Databricks


Audit Log Delivery Latency
Prod - records not delivered in 24 hours
Prod - 99 and 99.9 percentile of delivery duration
Staging - records not delivered in 24 hours
Staging - 99 and 99.9 percentile of delivery duration
Walldisplay::Trend
Daily Clusters Activities
EXIT
FIRE
databricks

The Unified Analytics Platform



PEOPLE → databricks → APPLICATIONS

Data Science

Data Engineering

Line of Business

DATABRICKS WORKSPACE
Interactive Data Science & Collaboration

DATABRI LOWS
Producti ow Automation

DATABRICKS

DATABRICKS I/O
Optimized Data
Access Layer

ark™

DB SERVERLESS
Fully Managed
Auto-Tuning Platform

DA ENTERPRISE SECURITY (DBES)
to-End Security & Compliance

Cloud Storage

Data Warehouses

Hadoop Storage

Deep Learning / ML

Streaming

Data Warehousing

and many others…

R is a first-class citizen

databricks

Databricks R Notebooks

- Notebooks are the cornerstone of Databricks workspace
- A notebook can attach to a cluster
- Users can mix languages in notebooks: R, Python, Scala, SQL, sh
- Markdown and visualizations are first-class elements
- R Namespace is configured with Spark API
- Jobs & dashboards are built on top of Notebooks

Scalability

- Databricks clusters run optimized Apache Spark
- R Notebooks support two popular R packages to program Spark
  - SparkR
    - R package distributed with Apache Spark
    - Exposes Spark DataFrames and several convenience methods in R
  - sparklyr
    - Spark backend for the popular dplyr package
    - Extensible API for other R packages to use Apache Spark

Spark and R together

Both SparkR and sparklyr

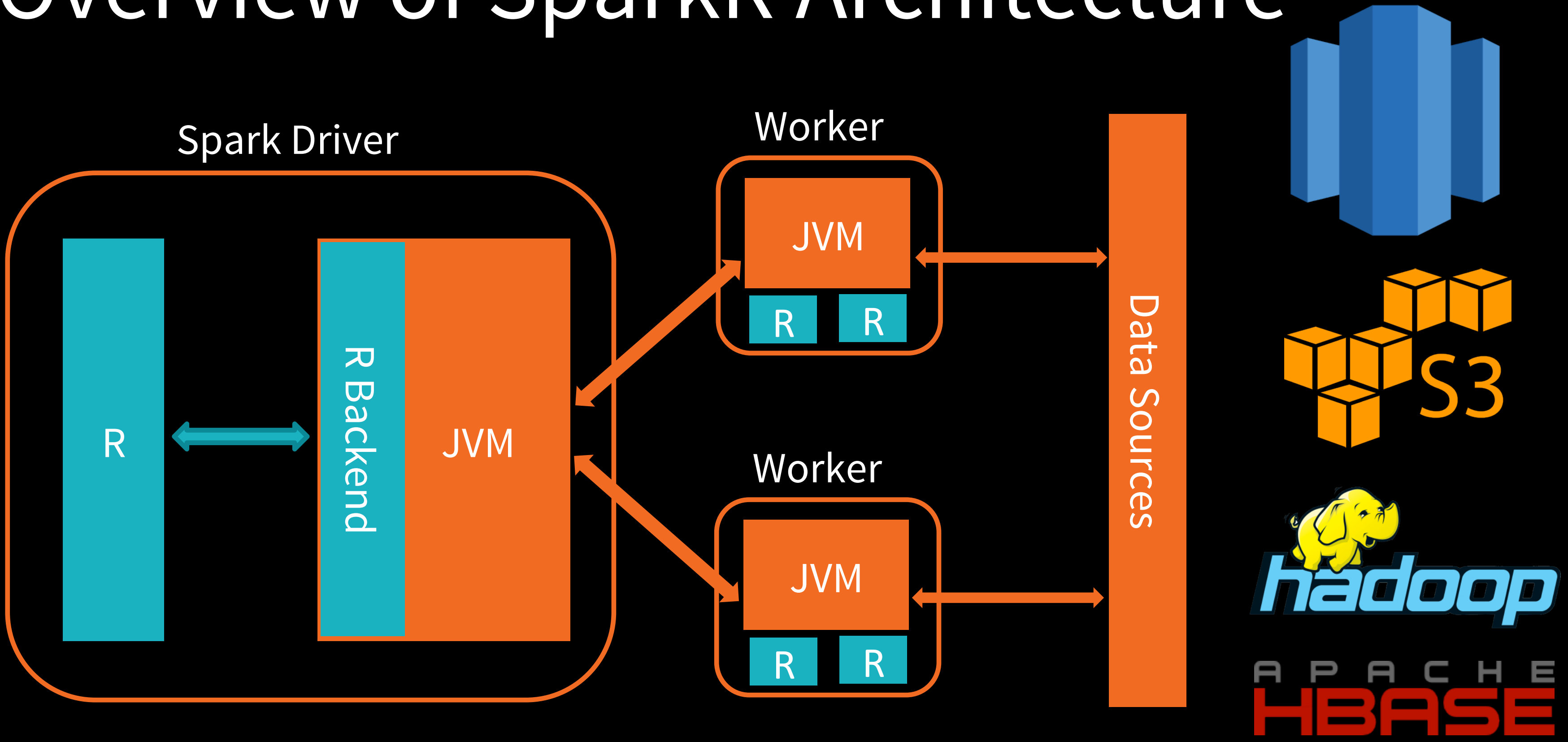
- Provides R front-end to Apache Spark
- Exposes Spark DataFrames (inspired by R & Pandas)
- Convenient interoperability between R and Spark DataFrames

robust distributed processing, data source, off-memory data

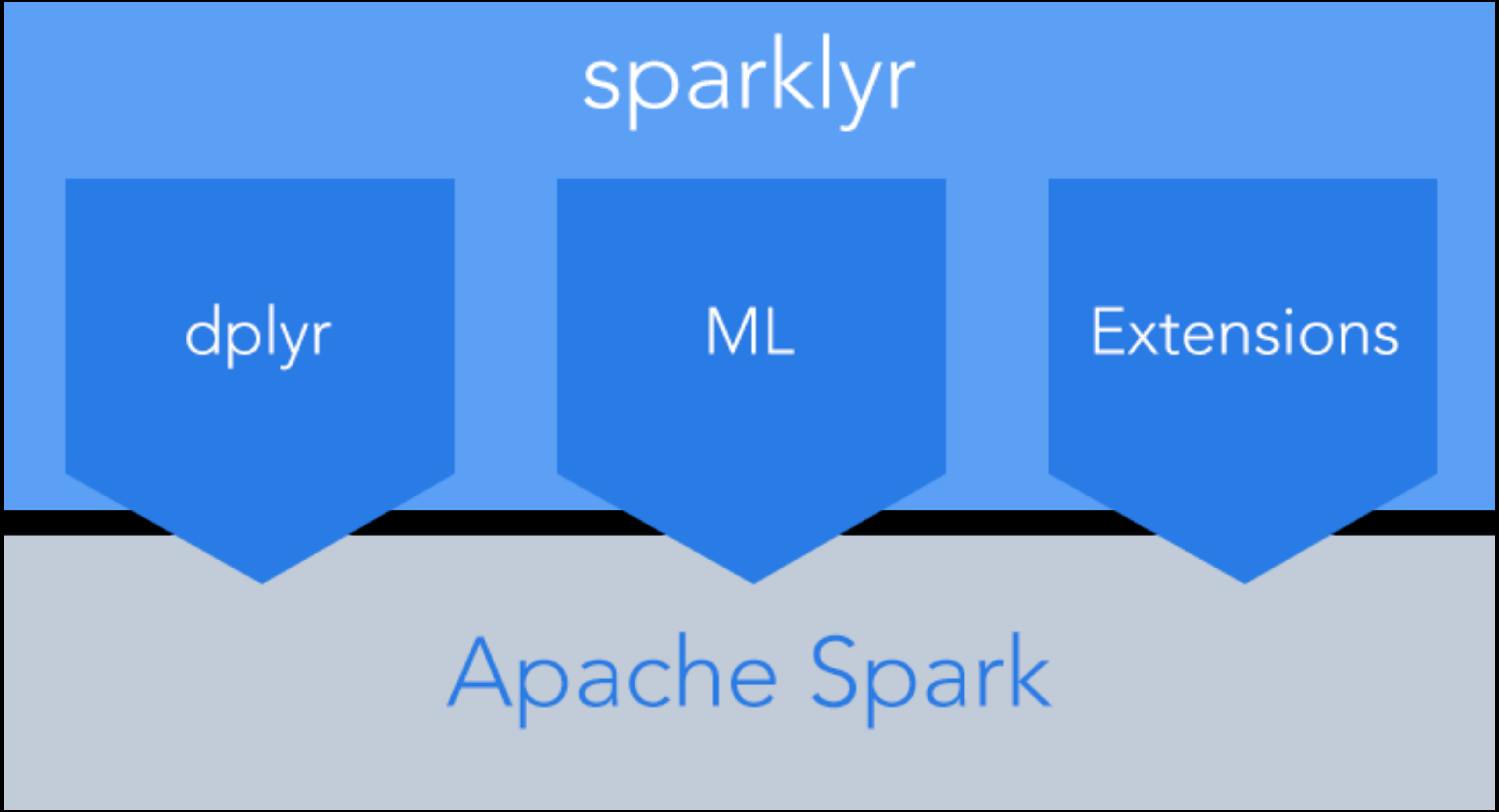
+

dynamic environment, interactivity, +10K packages, visualizations

Overview of SparkR Architecture



sparklyr stack



sparklyr

dplyr

ML

Extensions

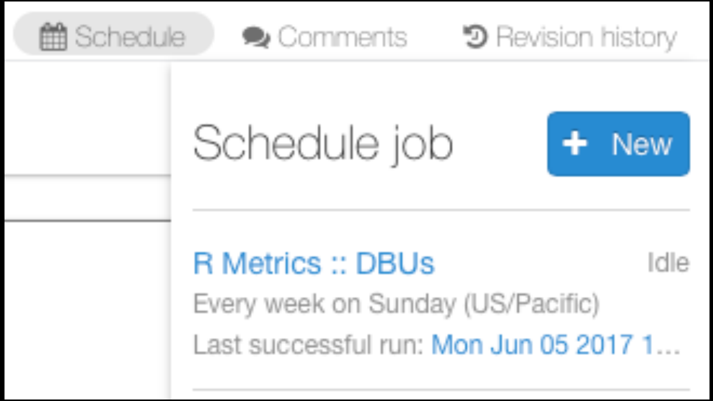
Apache Spark

Accessing (big) data

- Data is either stored on distributed file system or is streamed in
- At Databricks SparkR API is used to:
  - Read data using any of the existing 50 Spark Data Sources
    - Check out http://spark-pakcages.org
  - Ingest streaming data into Streaming SparkDataFrame
    - Checkout SSR: Structured Streaming on R for Machine Learning talk at Spark Summit

Reproducibility

- Notebooks are taking over the data field
  - Markdown, code and results live together
- Databricks (R) Notebooks:
  - Your version control system
  - Databricks jobs scheduling
- You can control all the elements of the environment:
  - Notebook version
  - Runtime: Spark + package versions

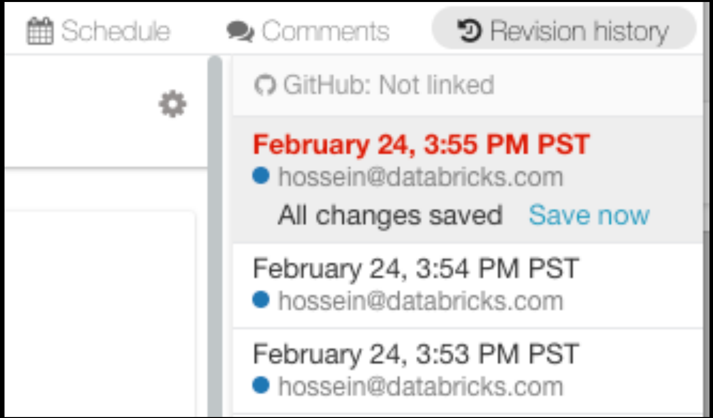


Schedule job

R Metrics :: DBUs Idle

Every week on Sunday (US/Pacific)

Last successful run: Mon Jun 05 2017 1...



GitHub: Not linked

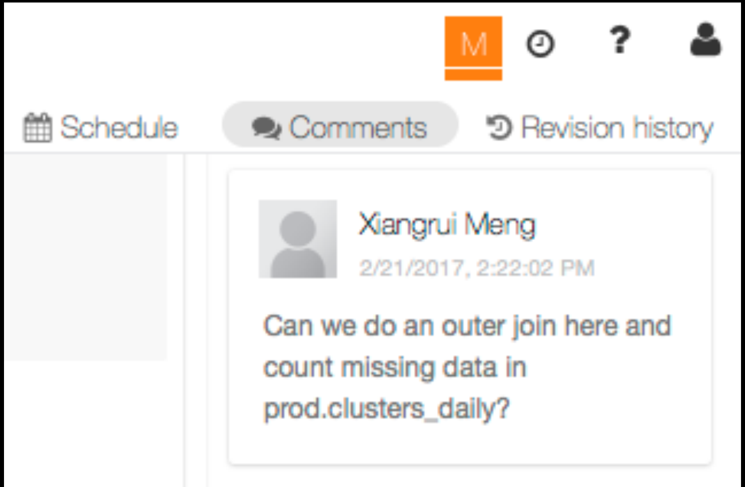
February 24, 3:55 PM PST
● hossein@databricks.com
All changes saved Save now

February 24, 3:54 PM PST
● hossein@databricks.com

February 24, 3:53 PM PST
● hossein@databricks.com

Collaboration

- Multiple users can simultaneously edit and run commands in a notebook:
  - Presence markers help uses with editing
  - Commenting help communication
  - Automatic snapshots to revert changes



Xiangrui Meng
2/21/2017, 2:22:02 PM

Can we do an outer join here and count missing data in prod.clusters_daily?

Sharing & publishing

- Dashboards are views on top of notebooks
  - user can build multiple dashboards from a single notebook
- Interactive dashboards using widgets
- Dashboard views of a job result can be shared and posted on wall displays
- Access control can restrict broader audience from editing/ running

Existing enterprise requirements

Security

- Authentication & authorization
- Data security & encryption
- Compliance
- Single Sign-on
- OpSec & access controls
- Compliance & auditing

Operations

- Resource management
- User management
- Monitoring
- Package management
- Version control

Deploying models built in R (coming soon)

Two simple steps for model scoring

1. SparkR models can be serialized and stored through API
2. Use a Databricks provided JAR in production to score score new data

More details soon …

Other enterprise use cases

- Running distributed Monte Carlo simulation
- Genomics
  - Using SparkR for sequencing alignment
  - predicting chemical structure & activity (Chemo-informatics)
  - Genotype and phenotype association to identify genomic variants and functional impact
- Modeling premium and pricing structure in insurance
- IOT device data analysis for commercial operations and marketing

Other interesting talks on Spark & R

Several talks on SparkR and sparklyr

All videos and slides will be available online

SPARK SUMMIT 2017
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 5 – 7, 2017 | SAN FRANCISCO
ORGANIZED BY databricks

Try Apache Spark in Databricks

UNIFIED ANALYTICS PLATFORM

Free (community) edition: https://community.cloud.databricks.com/

DATABRICKS RUNTIME 3.0

Apache Spark – optimized for the cloud

Thank You

Hossein Falaki @mhfalaki

databricks