Processing Large Data in R Using Apache Spark

Hossein Falaki
June 2017
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

- **Cluster 1**: service 1, service 2, service x, log-daemon
- **Cluster 2**: service y
- **Cluster 0**: service 0, service x, log-daemon

- **Amazon Kinesis** connects to:
  - Customer 1
  - Customer 2
  - Customer 3

- **DBFS**
  - **Raw record batch**
  - **Tables (parquet)**

- **Databricks Jobs**
  - Real-time analysis
  - Sync daemon
  - ETL jobs
  - Data analysis
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
The Unified Analytics Platform

**DATABRICKS WORKSPACE**
Interactive Data Science & Collaboration

**DATABRICKS RUNTIME**
DB SERVERLESS
Fully Managed Auto-Tuning Platform

**DATABRICKS I/O**
Optimized Data Access Layer

**DATABRICKS ENTERPRISE SECURITY (DBES)**
End-to-End Security & Compliance

**APPLICATIONS**

- Cloud Storage
- Data Warehouses
- Hadoop Storage

R is a first-class citizen

Data Science
Data Engineering
Line of Business

Deep Learning / ML
Streaming
Data Warehousing
and many others…
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

- 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-packages.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
Collaboration

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