Spark in the Hands of Data Scientists

Neelesh Srinivas Salian

September 2017
Algorithms Team
Data Science at
Stitch Fix
Who are they?

Data driven decision enablers
Planning and Optimization of different parts of the Business

- Verticals within Algorithms
- Cross-functional collaboration
- Providing insight and predictions for better decisions
Data Platform at Stitch Fix
What do they do?

Building and Maintaining the Algorithms Platform
Main Customers are Data Scientists

- Tools and Infrastructure
  - Data Warehouse Infrastructure
  - ETL Execution and Workflow
  - Improving the User Experience
  - Visualization
Lessons & Challenges
Lessons and Challenges

Usability is Key

- Bringing SQL into Spark
- Training Data Scientists’ with Spark basics
- Assistance in Tuning recommendations to optimize
- Maintain latest Spark releases
Data Architecture
Data Architecture - Highlights

Amazon S3 is the source of truth - stores tables
Hive Metastore - used to maintain schemas

- Spark - simple, complex pipelines
- Presto - adhoc queries
Spark at Stitch Fix
Why Spark?

Resourceful Framework
Easy to use and setup

- Different APIs - Python, Scala, Java

- Works with other parts of the architecture
  - Hive Metastore
  - S3

- Varied use cases
Spark Architecture
Genie
Genie

- Redirection Layer
- From Netflix
- Open Source - 2.x
- Used for Spark jobs - Load balance, Maintain Versions, Maintaining Multi-Tenant environment
Genie Data Model - Applications

Custom Builds of Spark

Spark 2.x may refer 2.1.0 or 2.2.0
Genie Data Model - Commands

Equivalent of *spark-submit* for each version

Tags (aliases)

Test commands with different configurations

Commands
Genie Data Model - Clusters

Red/ Black Cluster deployment

No disruption to Data Scientists’ workflow
Genie Data Model

- Commands
- Clusters
- Applications
- Jobs
Babylon
Sheriff of Babylon

- Job Submissions
- Genie client library
- Overrides to Spark default values
- Run Commands
  1. `run_spark`
  2. `run_sql`
  3. `run_with_json`
Additional Tools

- Custom SparkContext
- Metastore Library
- Table Statistics
- Python Query API
- Diagnostic Tool
Spaceman - Centralized Spark History Server
Spaceman
Archiving Log Data from all clusters

Singular entity to access Spark logs
Features

- Runs for each cluster - easy to add new clusters
- Checks for completed applications at intervals
- Stores History of Apps from all clusters for a month
- More Features to come
Thank you