85%
Getting your R Project into Production using Databricks

Michael Johnson
@MikeJohnsonZA
Data Science Lifecycle
Google File System (GFS) Predecessor to HDFS

2003

MapReduce 2004

Apache Hadoop Project

2006

Spark 2012

History of Azure Databricks
Google File System (GFS) 2003
Predecessor to HDFS

MapReduce 2004

Apache Hadoop Project 2006

Spark 2012

Databricks 2013

History of Azure Databricks
Architecture
Resilient Distributed Datasets

Directed Acyclic Graph
Spark Core

Resilient Distributed Datasets (RDDs)

Core API

Analysis APIs

SQL
Dataframes
Datasets
Spark SQL
ML lib
Spark ML
ML Pipelines

Language APIs

Java, Scala, Python
SparkR

Implementation

aws

Azure
One Environment for the Whole Team
One Click Setup
Scale
Security
Interactive Notebooks
R Studio Integration