Leveraging Apache Spark for IBM Machine Learning for z/OS

Eberhard Hechler
Executive Architect
Member Academy of Technology Leadership Team
IBM Analytics Platform
IBM Germany R&D Lab
Disclaimer

© Copyright IBM Corporation 2016. All rights reserved.
U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM, the IBM logo, ibm.com, DB2, and DB2 for z/OS are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml Other company, product, or service names may be trademarks or service marks of others.
Topics

- Introduction
- What is Apache Spark?
- Spark on z/OS
- Machine Learning for z/OS
  - Cognitive Assistant for Data Scientists
  - Hyper Parameter Optimization
  - Monitoring
- Coexistence DB2 Analytics Accelerator and ML for z/OS
- Summary
What is Apache Spark?
What is Apache Spark?

- Addressing **limitations** of Hadoop MapReduce programming model
  - No iterative programming, latency issues, ... 
- Using a **fault-tolerant** abstraction for **in-memory** cluster computing
  - Resilient Distributed Datasets (RDDs)
- Can be deployed on different **cluster managers**
  - YARN, MESOS, standalone 
- Supports a number of **languages**
  - Java, Scala, Python, SQL, R 
- Comes with a variety of **specialized libraries**
  - SQL, ML, Streaming, Graph 
- Enables additional use cases, user roles, tasks, e.g. 
  - **Data scientist** tasks: developing analytical models
  - **Using languages other than Cobol or Java**: R, Scala ...
Apache Spark is emerging as a key disruptive technology.

Expect more innovations to emerge

**Origin**

**Founding Sponsors:** Google, Amazon, SAP, IBM

**Sponsors:** Adobe, Apple, Bosch, Cisco, Cray, EMC, Ericsson, Facebook, Huawei, Informatica, Intel, Microsoft, Netapp, Pivotal, VMware.

**Affiliates:** many

**Aggressive Vision**

1. Unified Platform for Big Data Apps

   - Batch
   - Interactive
   - Streaming

   **Spark**

   - Hadoop
   - Cassandra
   - Mesos
   - Cloud Providers

   Uniform API for diverse workloads over diverse storage systems and runtimes

2. Standard Library for Big Data

   Big data apps lack libraries of common algorithms

   Spark’s generality + support for multiple languages make it suitable to offer this

   Python  Scala  Java  R

   SQL  ML  graph  ...

   Core
Resilient Distributed Dataset (RDD)

- **Key idea**: write programs in terms of transformations on distributed datasets
- RDDs are **immutable**
  - Modifications create new RDDs
- Holds **references to partition** objects
- Each partition is a subset of the overall data
- Partitions are **assigned to nodes** on the cluster
- Partitions are in **memory** by default
- RDDs keep information on their **lineage**
Spark Programming Model

- Operations on RDDs (datasets)
  - Transformation
  - Action

- Transformations use lazy evaluation
  - Executed only if an action requires it

- An application consist of a Directed Acyclic Graph (DAG)
  - Each action results in a separate batch job
  - Parallelism is determined by the number of RDD partitions
What is Apache Spark?

Languages

Spark Libraries

Spark Core

Cluster Manager

Data Abstraction

Languages: Java / Python / Scala / R

Spark Libraries: Spark SQL (Relational Operators), Spark MLLib (Machine Learning), Spark GraphX (Graph Processing), Spark Streaming (Real-Time Streaming)

Spark Core: General Execution Engine

Cluster Manager: YARN, MESOS, Standalone

Data Abstraction: HDFS / Cassandra / HBase / Parquet / ...
# RDDs / DataFrames / Datasets

<table>
<thead>
<tr>
<th>RDD</th>
<th>DataFrame</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spark Version</td>
<td>Spark 1.0</td>
<td>Spark 1.3</td>
</tr>
<tr>
<td>Released</td>
<td>May 2014</td>
<td>March 2015</td>
</tr>
<tr>
<td>API</td>
<td>RDD API</td>
<td>DataFrame API</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Functional programming, arbitrary data types, no optimized execution plan</td>
<td>Structured binary data (Tungsten(^1)), high level relational operation, Catalyst optimization (optimized execution plan)</td>
</tr>
<tr>
<td>Advantages</td>
<td>Strong typing, ability to use powerful lambda functions</td>
<td>Less code, better performance</td>
</tr>
</tbody>
</table>

1) Tungsten: fats in-memory encoding  
2) Sources at: [http://spark-packages.org/](http://spark-packages.org/)
Spark on z/OS
IBM Analytics Platform

IBM z/OS Platform for Apache Spark (Spark on z/OS)
Available since December 2015 via Open Source

Securely Integrate OLTP and Business Critical Data

Unique capability:
- Only found on Apache Spark on z/OS

Integrate:
- DB2 for z/OS, IMS, VSAM, PDSE, Syslog, SMF, ...
- Remote (non-z) data on distributed servers, Hadoop, Oracle, ...
- Defining security authorizations for instance using RACF
IBM Analytics Platform

IBM z/OS Platform for Apache Spark – Security Considerations

- Avoiding costly, cumbersome, and less secure z/OS data movement patterns
- Using Security Optimization Management (SOM) to cache user authorization information for logon processing
- Defining security authorizations using RACF to grant users access to the DB2 for z/OS
- Governing all Spark memory structures that contain sensitive data with z/OS security capabilities
- Using very granular integration with System Authorization Facility (SAF) interfaces for security configuration to suit needs
- Providing resource protection via Data Service server (AZK)
  - RACF classes
  - Top Secret classes
  - ACF2 (Access Control Facility) security
IBM Machine Learning for z/OS
IBM has transformed Machine Learning to Learning Machines

Easily create and deploy models in less time
Simplify model management
Continuously and automatically retrain models
Predict with more certainty

Faster time to value across more business initiatives
Higher quality predictions and recommendations
Optimized allocation of human resources
More impactful decision making
IBM Machine Learning (ML) for z/OS

- Consistent IBM offering
- On premise machine learning solution for z/OS
- Uses the same tools and technologies as the IBM Machine Learning (ML) Service
- Holistic approach, which offer data science experience (data prep, training and evaluation), model deployment, scoring, monitoring, and model retraining
- Can be used with and benefits from DB2 Analytics Accelerator
- Leverages z/OS Platform for Apache Spark (Apache Spark on z/OS)
IBM ML for z/OS – The Machine Learning Workflow

IBM Machine Learning (ML) for z/OS

(Data Science Experience)

- Training and scoring on z using Spark on z/OS as the backend data processor
- You can train models best fit for their business with the data on mainframe
- You can deploy the ML models and perform online scoring within transaction on mainframe
## IBM Machine Learning for z/OS – Faster Time-to-Value

### Differentiating Value

<table>
<thead>
<tr>
<th>Create better models in less time</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Rapidly optimize the algorithm that best fits the data and business scenario</td>
<td>Cognitive Assistant for Data Scientists (CADS)</td>
</tr>
<tr>
<td>• Provide optimal parameters for any given model</td>
<td>Hyper Parameter Optimization (HPO)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simplify model creation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Wizards make it easy for users to create and train a model</td>
<td>DSX Pipeline User Interface</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Improve models over time</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Monitor model accuracy with feedback data and performance history</td>
<td>Continuous Monitoring and Feedback Loop</td>
</tr>
<tr>
<td>• Notification of model performance deterioration for more efficient retraining</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Easily integrate with existing tools and applications</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Ease collaboration across users (e.g., Data Scientists and App Developers)</td>
<td>Modern RESTful APIs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simplify model management</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Easily manage thousands of models in an enterprise environment</td>
<td>Single UI for Deployment</td>
</tr>
</tbody>
</table>
Advantages

- Federated analytics across multiple data environments
- Increased currency of data & insights reduce latency
- Reduce cost and complexity of moving all data
- Benefits from DB2 Analytics Accelerator
- Integration with enterprise business applications
- Modern and consistent analytic skill across heterogeneous environment
IBM Machine Learning (ML) for zOS – Architecture

IBM Analytics Platform

Application Cluster

- GIT Repository Service
- Ingestion Service
- Transformation Service
- Jupyter Kernel Gateway
- Apache Toree
- Feedback Service
- Pipeline Service
- Monitor Service

Spark on z/OS Cluster

- Ingestion Lib
- Transformation Lib
- Pipeline Lib
- CADS/HPO lib

Liberty for z/OS

- Scoring Service

Individual Akka Svc Deployment Arch

- Nginx Load Balancer

NoteBook

- Notebook UI
- Jupyter Server

Kubernetes
- Linux
- z/OS

Auth Service

- Service Metadata
- DB2 for z/OS

DB2 for z/OS

- DVS Connector

zLDAP
- RACF

DB2
- IMS
- VSAM
- SMF
CADS and HPO – What are we addressing?

- Objective is to bring automation into key areas of large-scale data analysis and data scientist tasks
  - Overcome „analytic decision overload“ for Data Scientists
    - Different ML algorithms
    - Parameter settings
  - Enable Data Scientists to:
    - View and interact with decisions making process in an online fashion
    - Obtain rapid insights from data to answer key questions
  - Deliver a system for automated training of models for supervised ML tasks
  - Address the combinatorial explosion in choices of ML algorithms and implementations/platforms, their parameters and their compositions
CADS and HPO – Part of IBM ML for z/OS

- **Influencing factors** for training, optimization, continuous improvement to ensure accurateness of the classifier/learner:
  - Data selection, access, preparation, transformation …
  - The right ML model(s)
  - Hyper parameter setting and optimization
  - The ‘right’ adequate labeled test data, depending on the probability distribution of the available data
  - Choice and weight of ‘right’ columns of the tables

- **Cognitive Assistant for Data Scientists (CADS) with Hyper Parameter Optimization (HPO)**
  - Evaluation of various relevant Spark ML algorithms
  - Setting of corresponding hyper parameters
CADS and HPO – Some Parameter Examples

- For instances **parameter choices per algorithm** for the training algorithm (algorithm behavior) and the ML model (model capacity):
  - Kernel type
  - Learning rate
  - Number of layers in neural networks
  - Pruning strategy (decision trees)
    - Maximum depth of a decision tree
  - Number of trees in a random forest
  - . . .

Depends on model and labeled data sets
Model Evaluation and Selection

- Model selection and hyper parameter setting in ML for z/OS is based on defined measures:
  - Receiver Operating Characteristic (ROC) Curve
  - Precision Recall (PR) Curve
Monitoring Model Health by Evaluating Area under ROC or PR Curve
Monitoring Health by Evaluating Area under the PR Curve

Monitoring the Area under the ROC Curve

Monitoring the Area under the PR Curve
Impacting all Industries

**Finance/Banking**
Reduce customer churn while maximizing every marketing dollar

**Healthcare**
Improve pharmaceutical pricing and inventory planning based on claims patterns

**Airlines & Transportation**
Adjust passenger’s itineraries prior to delays and mechanical failures
## DB2 Analytics Accelerator & ML for z/OS – Complementing Values

<table>
<thead>
<tr>
<th>Situation #1</th>
<th>Situation #2</th>
<th>Situation #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML for z/OS</td>
<td>ML for z/OS</td>
<td>ML for z/OS</td>
</tr>
<tr>
<td></td>
<td>DB2 Analytics Accelerator</td>
<td>DB2 Analytics Accelerator</td>
</tr>
<tr>
<td>Small amount of z/OS data</td>
<td>Large amount of z/OS data</td>
<td>Large amount of z/OS data</td>
</tr>
<tr>
<td>zIPP / memory with</td>
<td>DataStage (or similar tool)</td>
<td>DataStage (or similar tool)</td>
</tr>
<tr>
<td>sufficient capacity</td>
<td>already used</td>
<td>already used</td>
</tr>
<tr>
<td>Scala, Python for data prep</td>
<td>SQL accepted for data prep</td>
<td>SQL accepted for data prep</td>
</tr>
<tr>
<td>required</td>
<td>Z capacity insufficient</td>
<td>Z capacity insufficient</td>
</tr>
<tr>
<td>Jupyter Notebook of ML for z/OS used</td>
<td>for data prep</td>
<td>for data prep</td>
</tr>
<tr>
<td>No SQL skills or SQL</td>
<td>ML for z/OS Jupyter notebook</td>
<td>ML for z/OS Jupyter notebook</td>
</tr>
<tr>
<td>not appropriate</td>
<td>used for addtl. data prep</td>
<td>used for addtl. data prep</td>
</tr>
<tr>
<td></td>
<td>(similar to SAS data cube</td>
<td>(similar to SAS data cube</td>
</tr>
<tr>
<td></td>
<td>build)</td>
<td>build)</td>
</tr>
<tr>
<td></td>
<td>Supporting broader</td>
<td>Supporting broader</td>
</tr>
<tr>
<td></td>
<td>data lake topologies</td>
<td>data lake topologies</td>
</tr>
<tr>
<td></td>
<td>Accommodating more</td>
<td>Accommodating more</td>
</tr>
<tr>
<td></td>
<td>data due to archiving</td>
<td>data due to archiving</td>
</tr>
<tr>
<td></td>
<td>Need for limited R support</td>
<td>Need for limited R support</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Situation #2**
- Large amount of z/OS data
- DataStage (or similar tool) already used
- SQL accepted for data prep
- Z capacity insufficient for data prep
- ML for z/OS Jupyter notebook used for addtl. data prep (similar to SAS data cube build)
- Supporting broader data lake topologies
- Accommodating more data due to archiving
- Need for limited R support
- ...

**Situation #3**
- Large amount of z/OS data
- DB2 Analytics Accelerator already deployed
- Addtl. points similar to situation #2
- Need for limited R support
- ...

**Situation #3**
- Large amount of z/OS data
- PMML needed
- Batch scoring
- SPSS Modeler and SPSS C&DS already used (integrates with the Accelerator)
- Interest in in-DB Analytics
- Need for limited R support
- Supporting broader data lake topologies
- ...
Notices and Disclaimers

- Copyright © 2016 by International Business Machines Corporation (IBM). No part of this document may be reproduced or transmitted in any form without written permission from IBM.

- U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY. IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

- Any statements regarding IBM’s future direction, intent or product plans are subject to change or withdrawal without notice.

- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

- It is the customer’s responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.
Notices and Disclaimers continued

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM’s products. IBM EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

- IBM, the IBM logo, ibm.com, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services®, Global Technology Services®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli®, Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.