SystemML: Declarative Machine Learning on Spark

05/03/19

Presented by: Juan Carrillo
Candidate for MASc. in Computer Software
Department of Electrical & Computer Engineering
University of Waterloo
Agenda

1. Introduction
2. SystemML core features
3. Experiments
4. Conclusions
5. Discussion
1. Introduction

Machine Learning for Big Data Analytics
1. Introduction

The problem, and the SystemML approach

**Usual workflow**
- Data Scientist
- Systems Programmer
- R or Python
- Scala
- Spark
- Results

**SystemML approach**
- Data Scientist
- R or Python
- SystemML
- Spark
- Results

- Time consuming
- Error prone
- Accelerates model development
- Simplifies deployment

Source: Spark Summit. Inside Apache SystemML
1. Introduction

SystemML background

- **Creation**: By researchers at the IBM Almaden Research Center
- **Open-source**: Spark Summit in San Francisco
- **Top Level Project**: Apache Software Foundation Board
- **Current release 1.2**: Deep learning functions, ultra-sparse data
SystemML core features
2. SystemML core features

Optimizer integration

Source: Spark Summit. Inside Apache SystemML
2. SystemML core features

Optimizer integration

Source: Spark Summit. Inside Apache SystemML
2. SystemML core features

Optimizer integration

Source: Spark Summit, Inside Apache SystemML
2. SystemML core features

Runtime integration

Distributed Matrix Representation

Buffer Pool Integration

Figure 2: Distributed Matrix Representation.

Figure 3: Buffer Pool Integration.
2. SystemML core features

Runtime integration

- Dynamic recompilation
  - Adapt the runtime plan to changing or initially unknown data characteristics

- Partitioning Operations
  - Partitioning-Preserving Operations
  - Partitioning-Exploiting Operations

- Specific Runtime Optimizations
  - Lazy Spark-Context Creation
  - Short-Circuit Read
  - Short-Circuit Collect
3. Experiments

End-to-End Performance

Figure 4: End-to-End Performance of Different Algorithms with Different Execution Modes.
3. Experiments

Runtime per Iteration

Figure 5: Runtime per Iteration of LinregCG and L2SVM with Different Execution Modes.
4. Conclusions

Takeaways and paper contributions

✓ Importance of DML as a high-level language to improve interoperability and scalability of Machine Learning models on Spark

✓ Multiple layers of abstraction and optimizations make SystemML a powerful tool for accelerating the development of Machine Learning models over Big Data

✓ Experimental evaluation on multiple ML models and datasets
Thanks for your attention
Discussion
5. Discussion

Research

1. Optimizer. How to optimize ML models over data streams?
2. Runtime. In dynamic recompilation, what could be unknown data characteristics?
3. Experiments. How SystemML might perform for the KNN algorithm?

Industry

5. Current capabilities compared to other tools such as Numpy, Scikit Learn, or TensorFlow?
6. Adoption in the current ML and Big Data user base?
7. SystemML in Cloud computing infrastructure. Beyond IBM?