SystemDS: Introduction and Overview

Rohit Lokwani
Sravan Hande
Tharun Kumar Reddy Karasani

Agenda

1. Introduction to SystemDS
2. Architecture Discussion
3. Installation and Integration
4. Performance Benchmarking
5. System Demo

SystemDS

SystemDS is an open source ML system for the end-to-end data science lifecycle from data integration, cleaning, and feature engineering, over efficient, local and distributed ML model training, to deployment and serving

1

Customizablitity

Algorithm customaziability via R-like and Python-like languages

2

Scalability

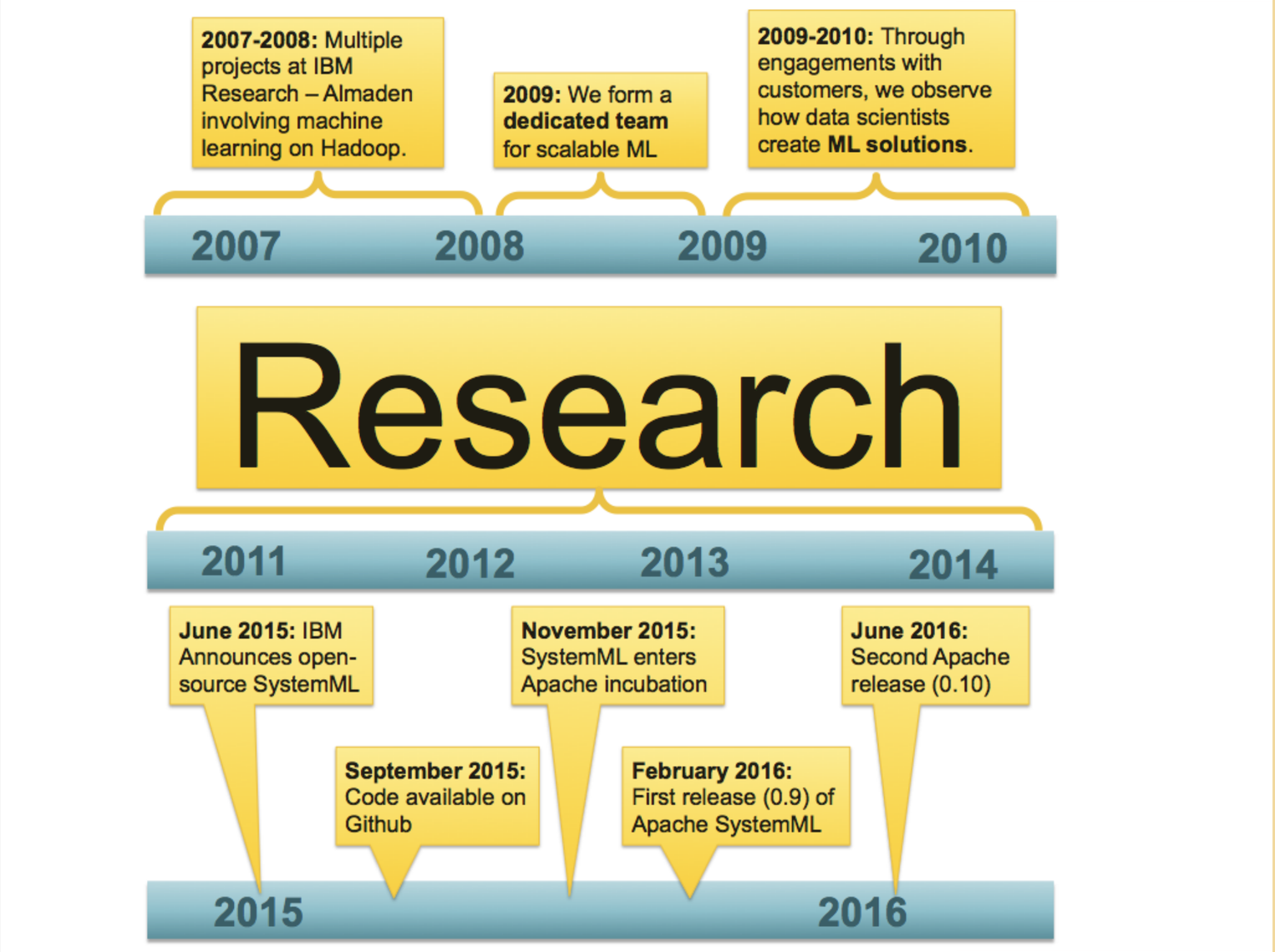
Multiple execution nodes including Spark MLContext, Batch, Standalone

3

Automatic Optimization

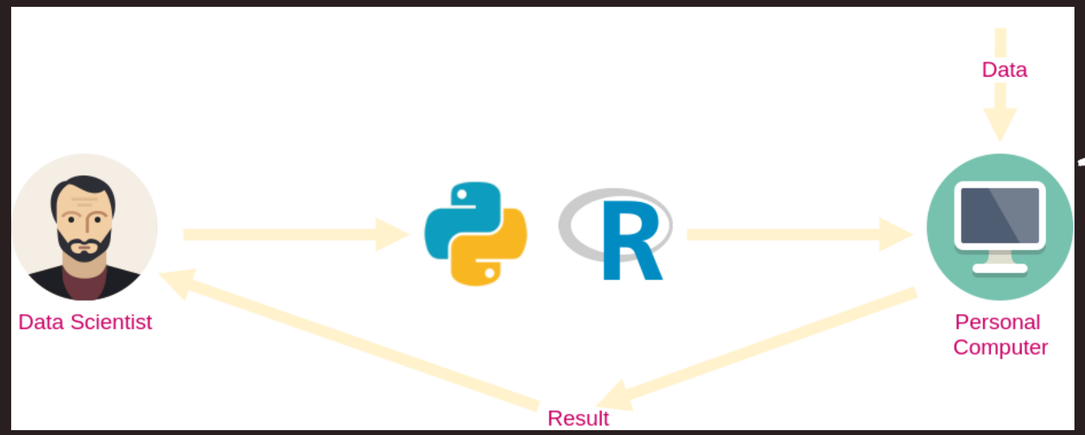
Based on data and cluster characteristics to ensure efficiency and scalability

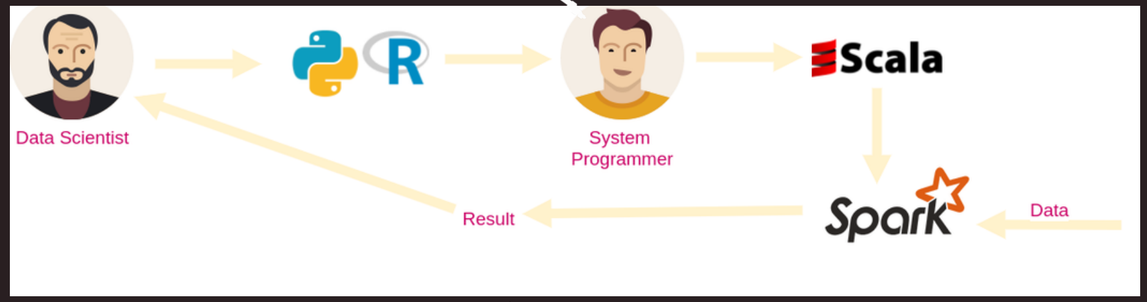
SystemDS Timeline



The History of Apache SystemML

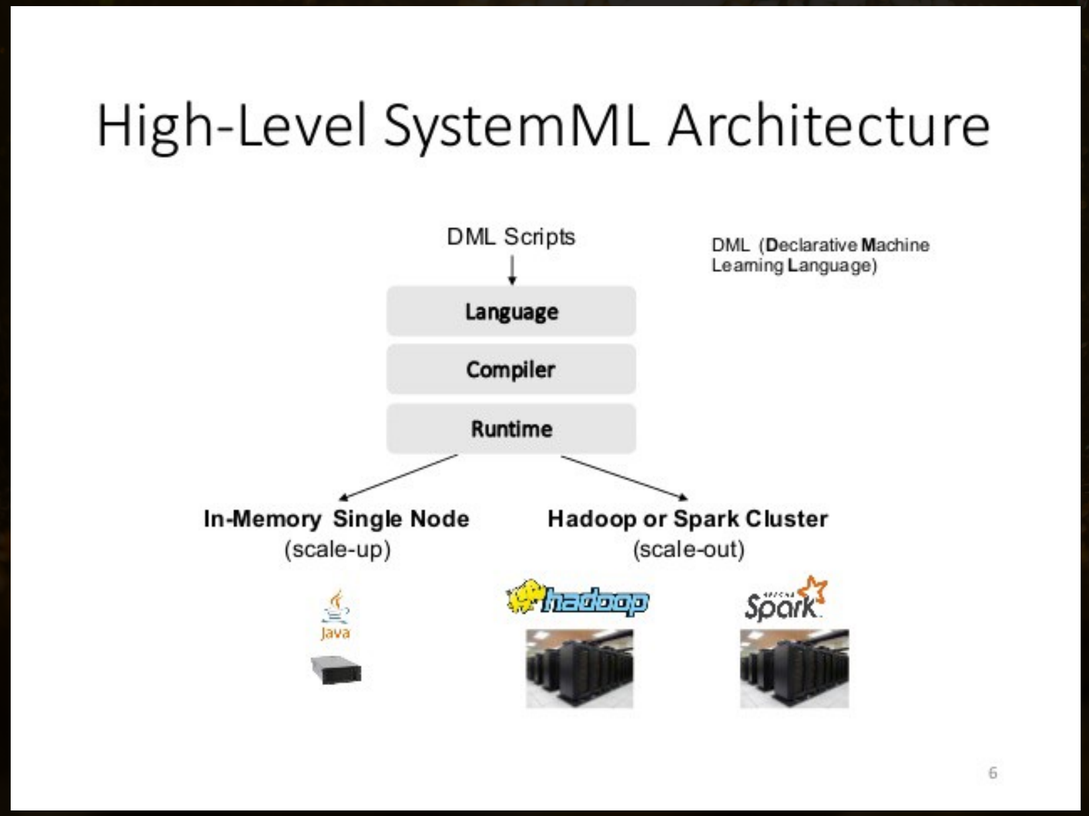
Efficiency and Scalability



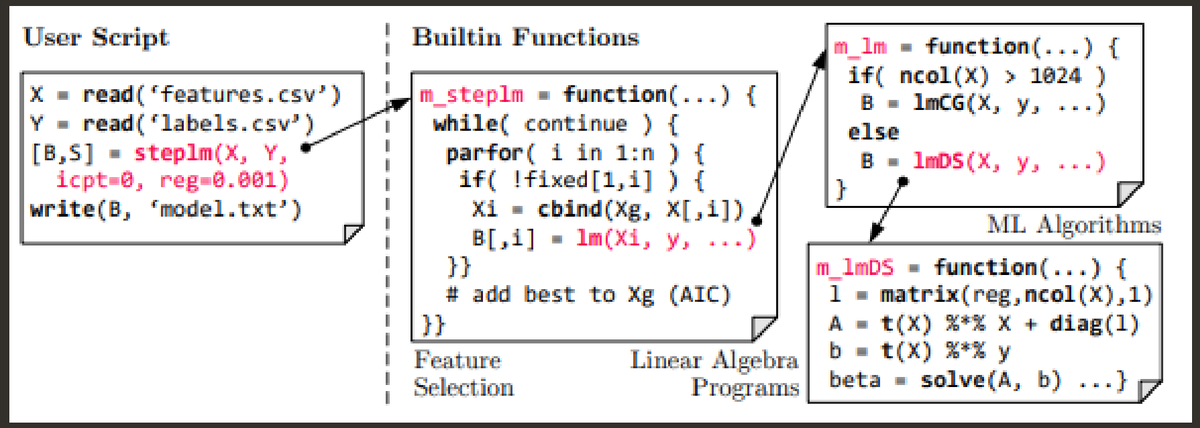


https://systemds.apache.org/

Platform Independence



Ease of programming



https://systemds.apache.org/

SystemDS + Spark Overview

Consistency

Immutable RDDs, use lineage graphs to recove

01

Scalability

Multiple execution nodes

02

Replication

Replicate RDDs with no or some transformations

03

Partitioning

Logical parititioning in RDDs

04

Ease of programming

Language flexibility, minimum LOC

05

Documentation

Lack of detailed documentation

06

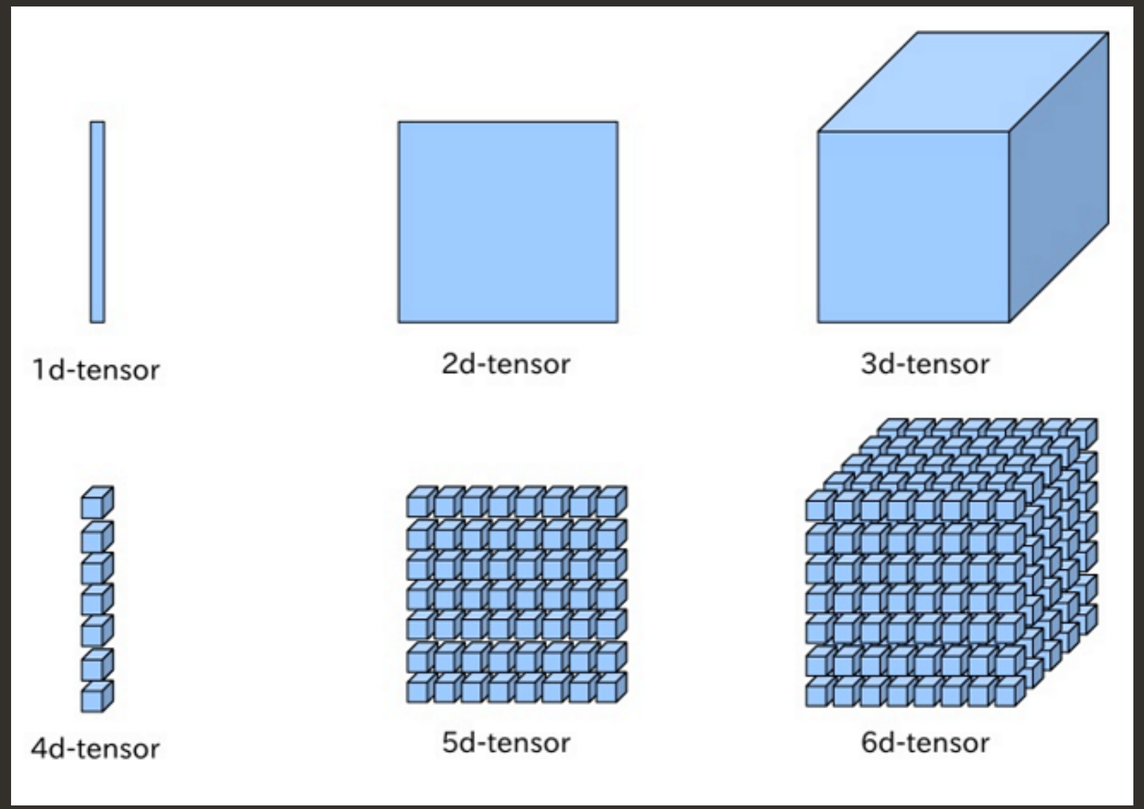
Things to know about Spark RDD(2022)

SystemDS on Spark

- Distributed processing
- Utilizes spark RDD
- In-memory Processing
- Write your topic or idea

Data Model

- Basic Tensor Blocks (CPU)
- Data Tensor Blocks (GPU)
- Distributed Tensor Blocks (Spark)
- Federated Tensors



User Interface

CLI and DML

Briefly elaborate on what you want to discuss.

Java API

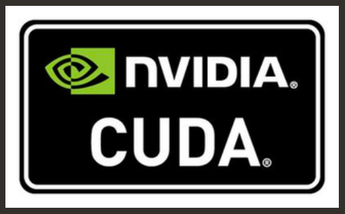
Briefly elaborate on what you want to discuss.

Python API

Briefly elaborate on what you want to discuss.

Installation

- Dependancies (Java8)
- Python & Jupyter
- SystemDS pip package
- Required GPU libraries



https://systemds.apache.org/docs/2.2.1/api/python/getting_started/install.html

Dataset - Airbnb

- 4 Cities
- Listings, Availability, Reviews

  and Neighbourhood data
- 350MB
- Data Analysis and Regression

Backend Support

Local

GPU

Distributed

Discuss about support for SPARK and other distributed frameworks

Federated

Discuss about federated environment

Benchmark Comparison

Baseline Comparison
TensorFlow2 Comparison

System Demo

Demo of SystemDS in distributed and federated environments

Conclusion / Summary

Add a little bit of body text

References

- https://systemds.apache.org/
- https://github.com/apache/systemds
- https://en.wikipedia.org/wiki/Apache_SystemDS
- https://apache.github.io/systemds/
- https://apache.github.io/systemds/site/algorithms-reference

Questions?

Thank You!