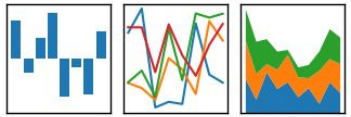
Data Science with Python

DataFrames in Pandas, Spark and GraphLab Create

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Spark

GraphLab

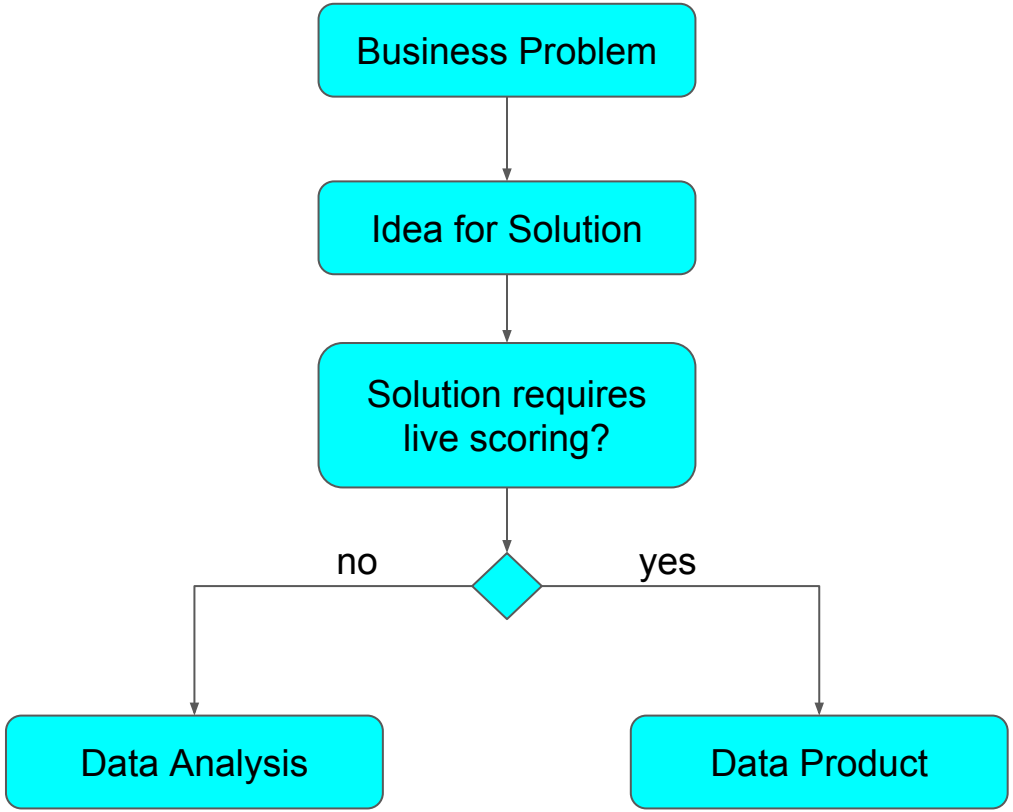
About Me

- Head of the Data Science dept. at Novomatic HQ
- Organizer of the Modern Data Science Tools Meetup (https://www.meetup.com/Vienna-Modern-Data-Science/)
- Background in Theoretical Physics and SW Engineering

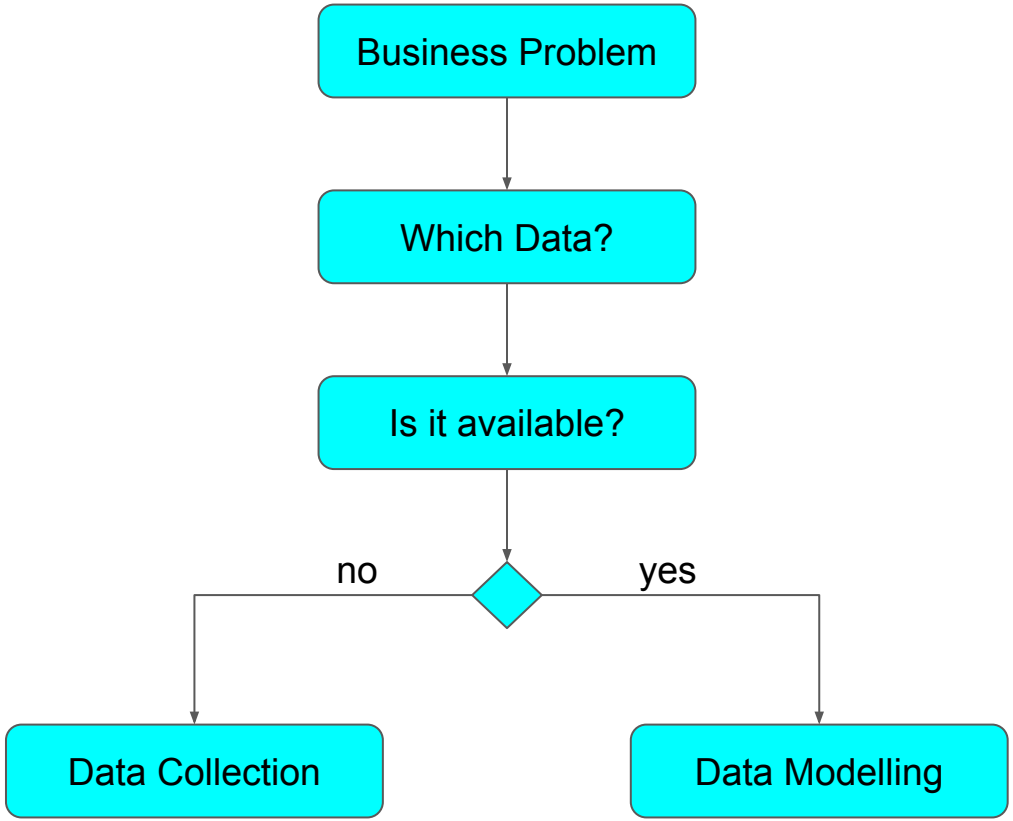
NOVOMATIC

The Data Science Process

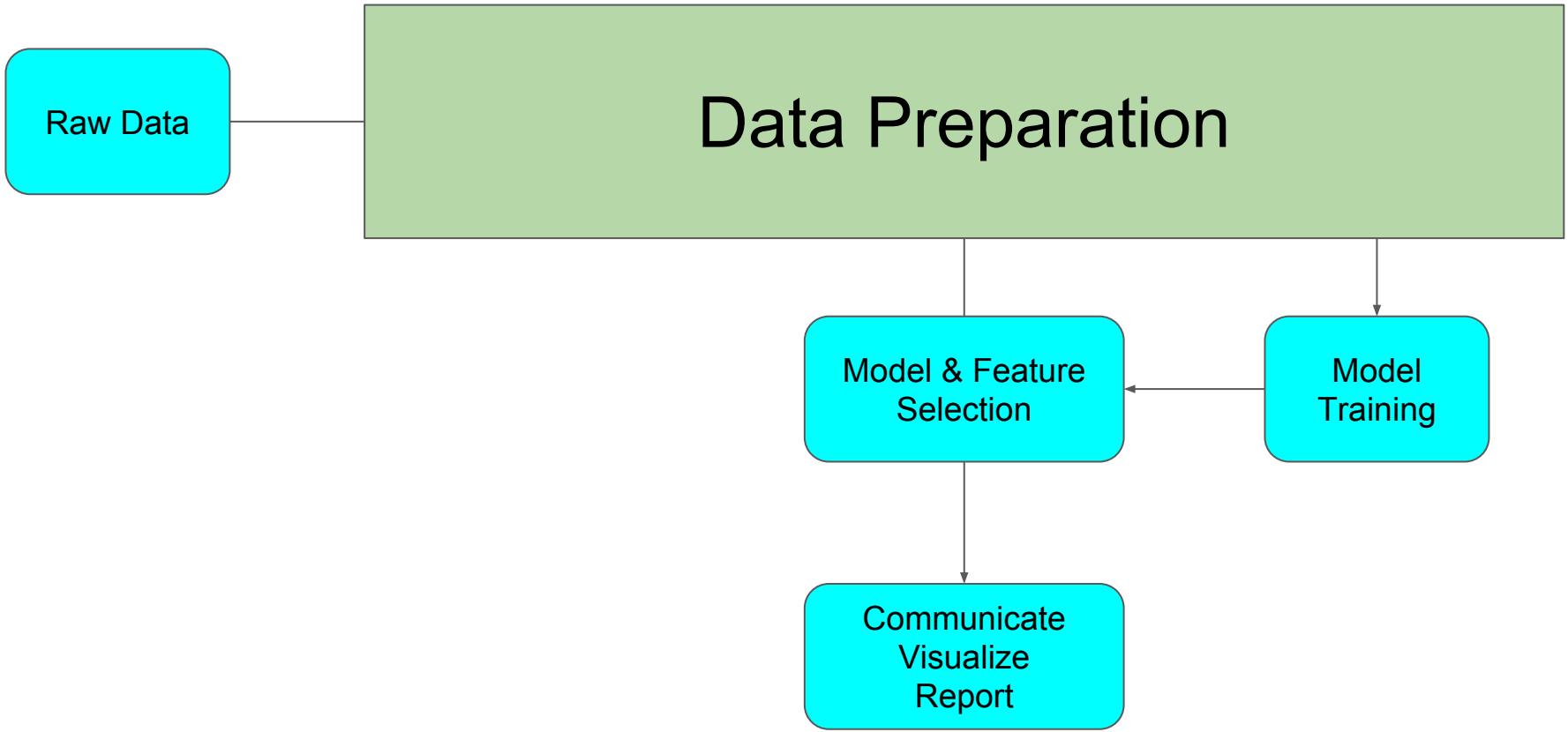
Which Type of Solution?



Which Data?


Business Problem
Which Data?
Is it available?
no
yes
Data Collection
Data Modelling

Data Analysis with Available Data



My DataFrame Journey

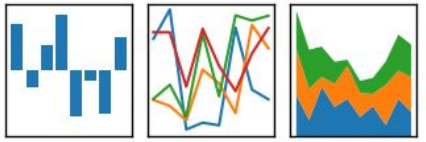
Recommender Systems

2		2	4	5	
5		4			1
		5		2	
	1		5		4
		4			2
4	5		1		

The Python Data Science Ecosystem

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



matplotlib

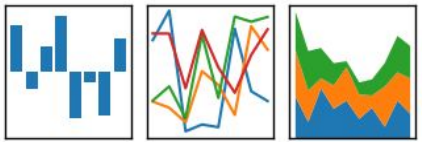
Evaluate RecSys Libraries

Problem: Spark MLlib only has Matrix Factorization

Solution:

≈ pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



+

A Day in the Life of a Data Scientist

Model Training

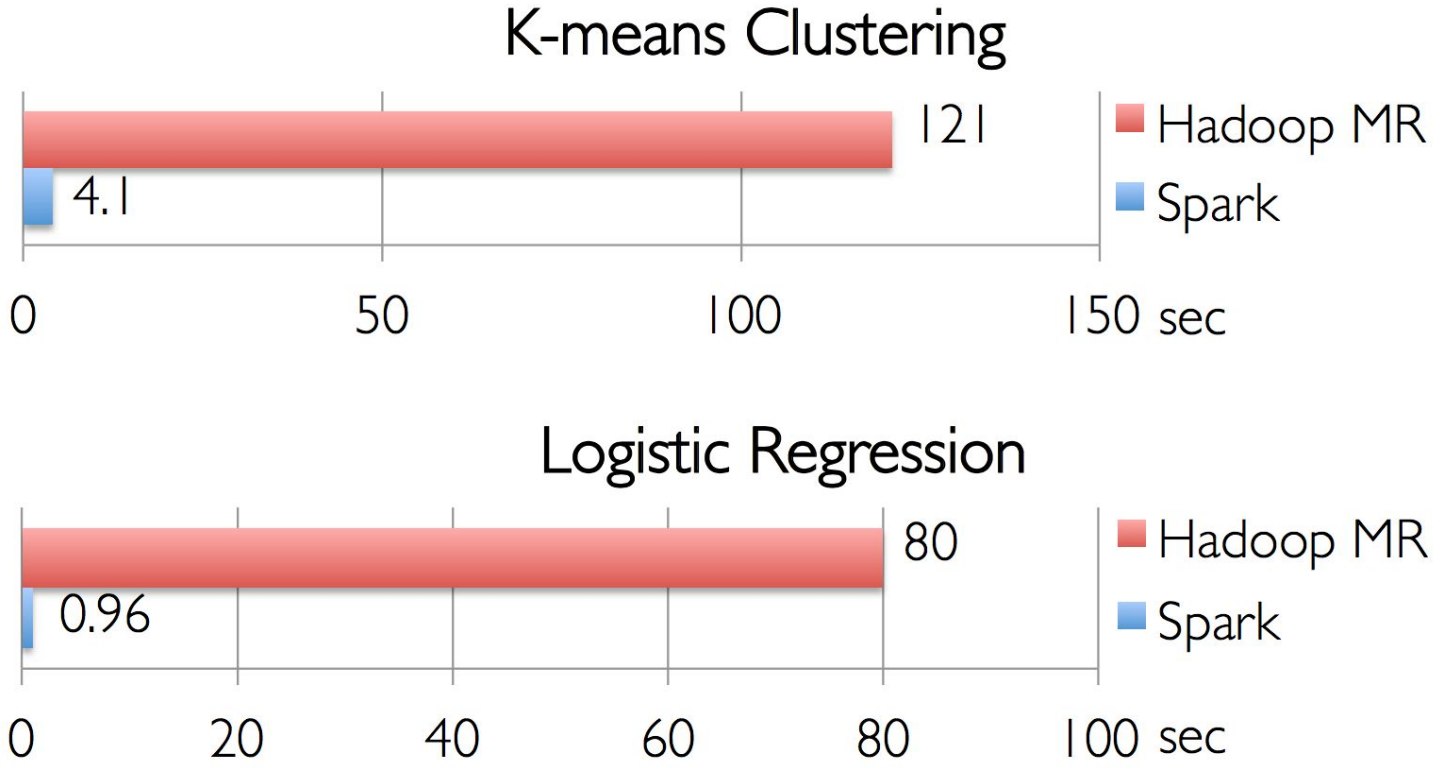
Data Preparation

What I need from a DataFrame library

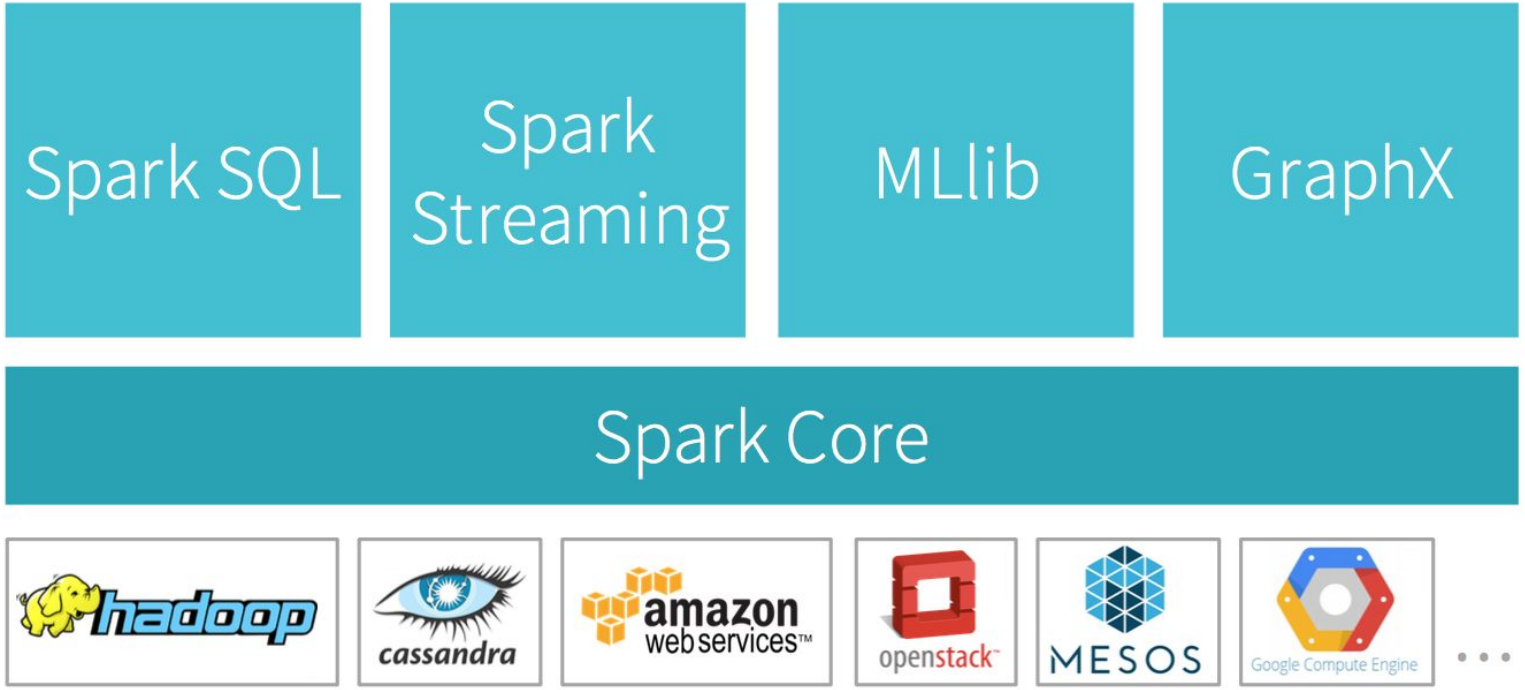
- Intuitive & easy-to-use API
- Good documentation
- Scalability
- Performance
- Free to use


Spark

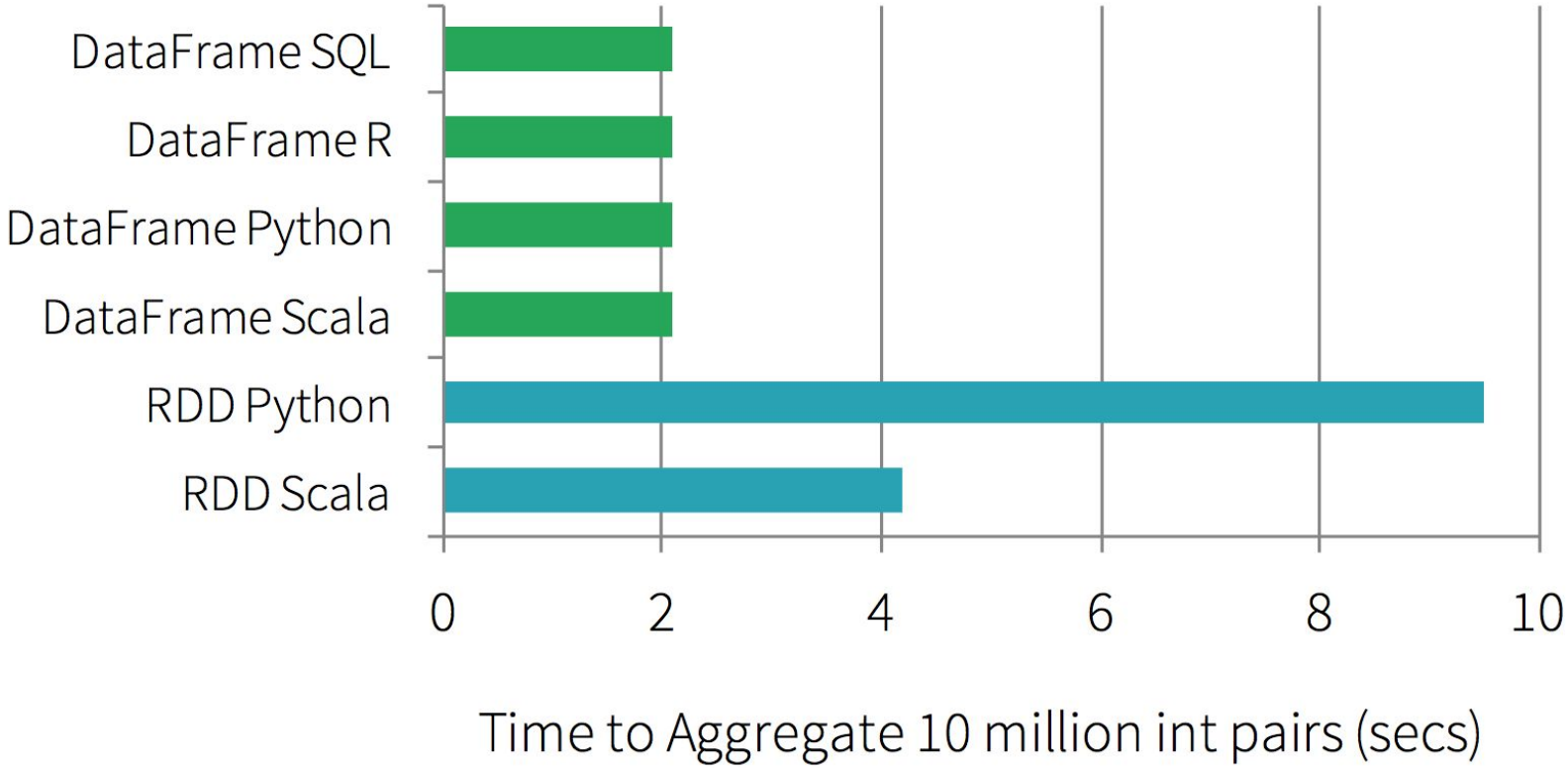
In-Memory Distributed Computing Engine for Big Data



Spark Architecture



RDDs, DataFrames, DataSets



Spark Resources

- Databricks: Spark as a Service
  - Very easy to spawn a Spark cluster
  - Databricks Community Edition
    - https://databricks.com/try-databricks

- MOOC on edX:
  - https://www.edx.org/xseries/data-science-engineering-apache-spark

Spark DataFrames

What I like:

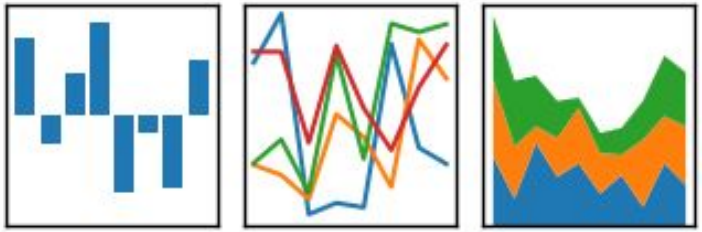
- Free to use / open source
- SQL-based API
- Big Data scalable
- Parquet
- Zeppelin Visualization
- Same API for Streaming

What I don’t like:

- Documentation
- Memory bottleneck
- Not possible to concatenate DataFrames horizontally
- Cannot apply UDF’s on aggregated fields (at least in Python)

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Pandas DataFrames

What I like:

- Free to use / open source
- Very mature library
- Pivot available (http://pbpython.com/pandas-pivot-table-explained.html)
- Parquet (NEW)
- Seaborn (https://web.stanford.edu/~mwaskom/software/seaborn/)
- Faster than GraphLab for small problems

What I don't like:

- API
  - Indexing
- Documentation
- Memory bottleneck
- No parallel processing / single-threaded


turi
create intelligence™

GraphLab Create

- ML library + SFrame API
  - Written in C++
  - Optimized disk access
  - Python API

- Turi was purchased by Apple in 2016
  - No further development and support for GraphLab Create and the SFrame projects
  - Free academic license:
    - https://turi.com/download/academic.html

- MOOC on Coursera:
  - https://www.coursera.org/specializations/machine-learning

GraphLab Create SFrames

What I like:

- API
- Documentation
- Very mature
- Parallel processing (single node)
- No memory bottleneck
- Pivot available (unpack/pack)

What I don't like:

- Not maintained any more

- SFrames are open source: https://github.com/turi-code/SFrame
- Sort your SFrame before saving it on disk

API Comparison

Summary

	Spark	Pandas	GraphLab Create
API	OK, excellent with SQL	OK	Excellent
Documentation	OK	OK	Excellent
Big Data	Excellent	No	No
Medium Data	OK	Possible	Excellent
Small Data	OK	Excellent	Excellent
Typed	Yes (in Scala)	No	No

Conclusions

- For Big Data problems use Spark DataFrames
- Please fork SFrame and start maintaining it
- For Small Data problems use Pandas

Demo

Thank you!

bpirvu@novomatic.com

NOVOMATIC