Pandas vs Koalas: The Ultimate Showdown

Amanda Moran
Solutions Architect, Databricks

Sadly, we won't be talking about this ...

But we will be talking about DS at Scale ...

... which is just as good!

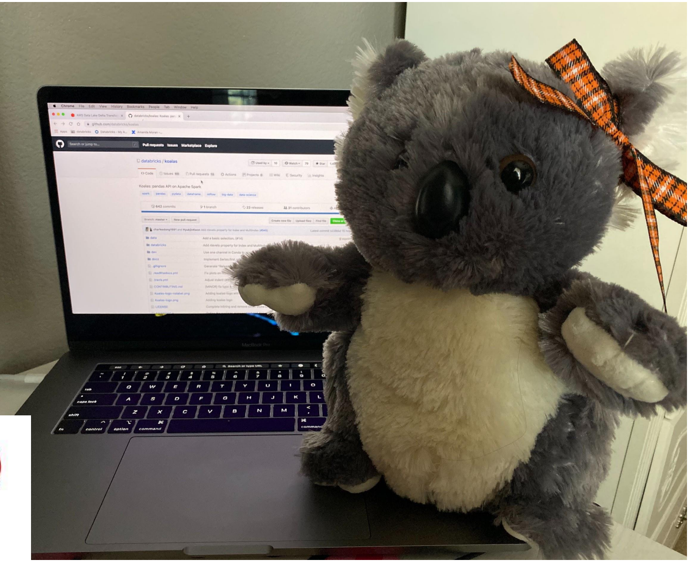
- Introduction to doing Data Science at Scale
- A few words on Pandas
- What is Apache Spark?
- What is Koalas?
- Demo
- Ultimate Showdown!

A Little about Amanda ...

- Solutions Architect @ Databricks
- MS Computer Science, BS Biology
- Previously: HP, Teradata, DataStax, Esgyn
- PMC and Apache Committer on Apache Trafodion
- 5 Different Distributed Systems
- Instructor for Udacity Data Engineering Nanodegree

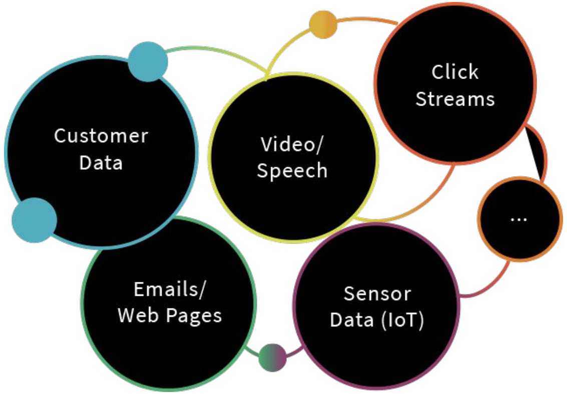
Let’s have a Contest!

- If you want to win this Koala:
  - Create a great live tweet on Twitter
  - Tag me **@AmandaK_Data**
  - Use the hashtag
    - **#pydataNYCKoalas**
    - **#pyspark**
- Stick around after the talk



Why Should a Data Scientist Care about Scale?

- Huge amounts of data from many sources
  - Click steam, Customers data, IOT, video/speech
  - And this isn't going away -- only growing
- Large data + simple algorithms = better models
- Documented by Google in 2009
  - White Paper: *The Reasonable Effectiveness of Data*



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

What is Pandas?

pandas

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

- Authored by Wes McKinney in 2008
- The standard Python tool for data manipulation/analysis
- Can deal with a lot of different situations, including:
  - Basic statistical analysis
  - Handling missing data
  - Time series, categorical variables, strings

Why Pandas?

Easy to start with Pandas

- Default choice for teaching
- Easy to install and use on any laptop
- Easy to write tests with all the python
- Huge community
- Enormous API for data manipulation
- Integration with visualization, ML tools

What is Apache Spark?

- Open Source
- De facto unified analytics engine for large-scale data processing
  - Streaming
  - ETL
  - Machine Learning
- PySpark API for Python
  - API support also for Scale, R, SQL

What's Wrong with PySpark?

- Nothing at all
- But integration isn't seamless
  - Both have dataframes

Pandas

- Standard for *single machine* workloads
- Small data

Apache Spark

- Standard for *distributed* workloads
- Big data

What’s Wrong with PySpark?

Pandas DataFrame vs Spark DataFrame

	pandas DataFrame	Spark DataFrame
Mutability	Mutable	Immutable

Value count	df[‘col’].value_counts()	df.groupBy(df[‘col’]).count()
		.orderBy(‘count’, ascending = False)

Pandas vs PySpark

Pandas

```
import pandas as pd
df =
pd.read_csv("my_data.csv")

df.columns = ['x', 'y', 'z1']

df['x2'] = df.x * df.x
```

PySpark

```
df = (spark.read
    .option("inferSchema", "true")
    .option("comment", True)
    .csv("my_data.csv"))

df = df.toDF(‘x’, ‘y’, ‘z1’)

df = df.withColumn(‘x2’, df.x*df.x)
```

What is Koalas?

- Announced April 24, 2019
- Pure Open Source Python library
- Aims at providing the pandas API on top of Apache Spark:
  - Unifies the two ecosystems with a familiar API
  - Seamless transition between small and large data

What is good about Koalas?

- Be immediately productive with Spark
  - No learning curve
- Have a single codebase that works both with pandas
  - Tests
  - Small datasets
  - Spark distributed datasets
  - Large datasets

Pandas vs Koalas

Pandas

```
import pandas as pd
df =
pd.read_csv("my_data.csv")

df.columns = ['x', 'y', 'z1']

df['x2'] = df.x * df.x
```

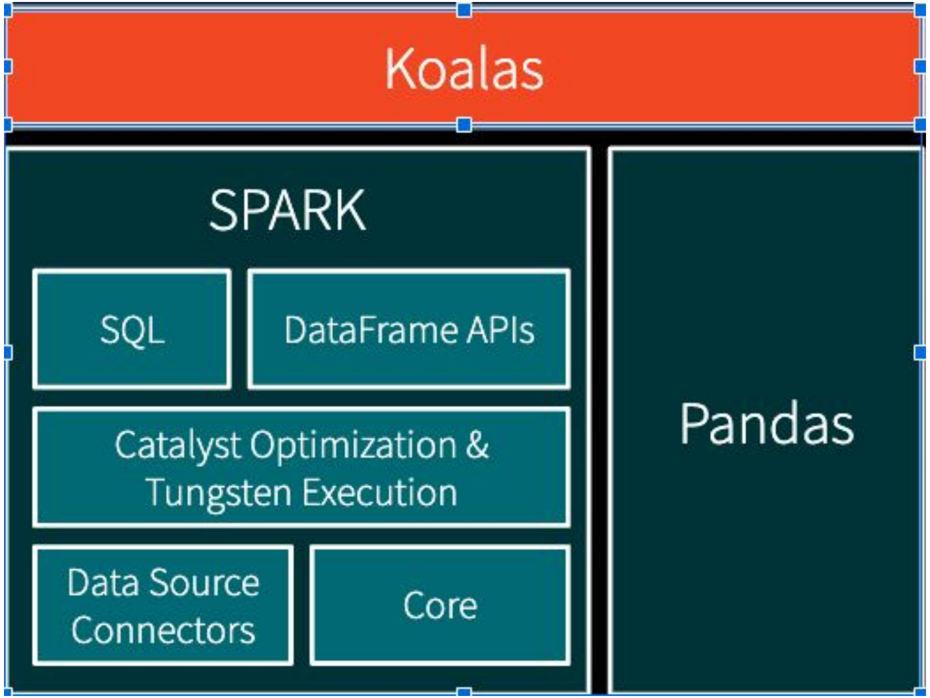
Koalas

```
import databricks.koalas as ks
df = ks.read_csv("my_data.csv")

df.columns = ['x', 'y', 'z1']

df['x2'] = df.x * df.x
```

Koalas Architecture



Koalas

SPARK

SQL

DataFrame APIs

Catalyst Optimization & Tungsten Execution

Data Source Connectors

Core

Pandas

Demo

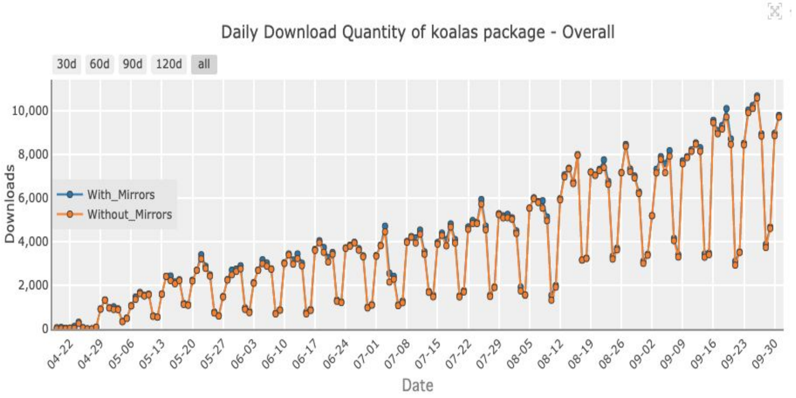
Ultimate Showdown: Who is the Winner?

- YOU!
  - More tools to be productive
- Koalas for scale
- pandas for learning and small data

Quickly Gaining Traction

- Bi-weekly releases!
- > 500 patches merged since announcement
- > 20 significant contributors outside of Databricks
- > 8k daily downloads

Downloads last day: 9,724
Downloads last week: 56,516
Downloads last month: 216,658



Status

- Bi-weekly releases, very active community with daily changes

- The most common functions have been implemented:
  - 60% of the DataFrame / Series API
  - 60% of the DataFrameGroupBy / SeriesGroupBy API
  - 15% of the Index / MultiIndex API
  - to_datetime, get_dummies, …

How to Get Started

- `pip install koalas`
- `conda install koalas`
  - More instructions on https://github.com/databricks/koalas
- Documentation
  - https://koalas.readthedocs.io/en/latest/
- Databricks Community Edition:
  - https://databricks.com/signup/signup-community

Get Involved!

- Contribute to the code
  - https://github.com/databricks/koalas

This talk:

- Notebooks and Slides
  - https://github.com/amandamoran/pydatanyc

SCAN ME

Thank you!