Monary

Really Fast Analysis with NumPy and MongoDB
Matt Cotter & Kyle Suarez

Imagine that you are a...

...data scientist

...biologist

...MongoDB user

...someone that has taken a taxi

...and you want to analyze data.

What if you only need an average?

If the data is stored in a...

❖ List of Python Dicts

~12 million numbers/sec : (

❖ Python List

110 million numbers/sec : |

❖ `numpy.ndarray`

500 million numbers/sec :)

NumPy

❖ <u>Num</u>erical <u>Py</u>thon

❖ Arrays and Matrices

❖ Lots of math

How do we get data from our database into NumPy?

Monary

About us

Kyle Suarez

🐦 @ksuarz

🐙 ksuarz

Matt Cotter

🐦 @MattWCotter

🐙 machyne

College seniors
Interns at MongoDB Inc.
Core contributors to Monary

Why MongoDB?

MongoDB: a NoSQL database

❖ open-source document database
❖ data in binary JSON format

```
{
    "names": ["Matt", "Kyle"],
    "ages": [21, 21],
    "talk": {"pygotham": "Monary"}
}
```

Distribution

❖ Scale

❖ Read Capacity

Data is unpredictable.

❖ Flexible schemas:

  ■ Accommodates varying data structures
  ■ Not all documents need to have the same fields

❖ Works well with NumPy's masked arrays
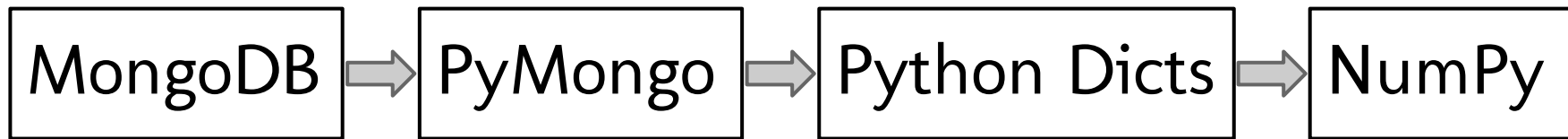
What happens in between?

MongoDB   ????   NumPy

PyMongo Workflow

MongoDB → PyMongo → Python Dicts → NumPy

This is pretty fast:

With the server and processes both running on this laptop...

About 150,000 documents read per second.

But...

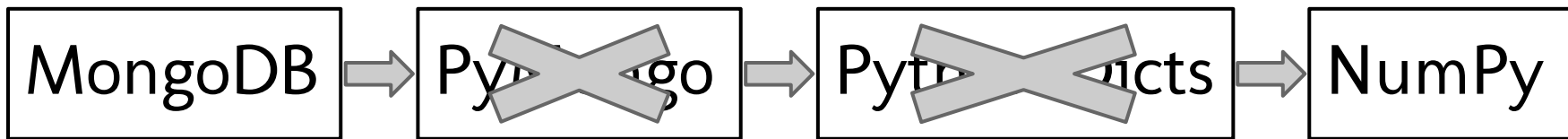❖ Object creation overhead is expensive.
❖ Dicts are only intermediate storage vessels.

Can we get faster?

Enter Monary.

MongoDB → PyMongo → Python Dicts → NumPy

Monary Workflow

MongoDB → Monary → NumPy

This is *really* fast!

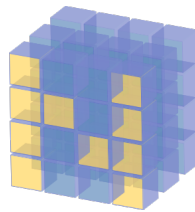Same test as before...

This time over 1,700,000 reads per second.

What is Monary?

Monary: the bridge between mongoDB & NumPy

Monary Overview

❖ author: David J. C. Beach

❖ operations: CRUD only

❖ dependencies: NumPy

❖ driver: MongoDB C Driver 0.98.0

How does it work?

- ❖ Implemented in C:
    - ■ MongoDB C driver
    - ■ MongoDB C BSON library

- ❖ NumPy + ctypes:
    - ■ Create NumPy arrays
    - ■ Pass pointers to the arrays into C

Demo

Monary 0.3.0 Documentation

Overview

Monary provides a Python interface for fast column queries from MongoDB. It is recommended over PyMongo when requiring large bulk reads from a database into NumPy arrays.

Note that Monary is still in beta. There are no guarantees of API stability; furthermore, dependencies may change in the future.

Monary is written by David J. C. Beach.

Installing / Upgrading
  Instructions on how to get the distribution.

Tutorial
  Getting started quickly with Monary.

Examples
  Examples of how to perform specific tasks.

Type Reference
  In-depth explanation of how to use various functions and DSON types.

Frequently Asked Questions
  Frequently asked questions about Monary.

See tinyurl.com/monary-docs

Dependencies

Monary depends on PyMongo and NumPy.

Monary uses the MongoDB C driver, which comes bundled as part of the module.

Issues

All issues can be reported by opening up an issue on the Monary BitBucket issues page.

Contributing

Monary is an open-source project and is hosted on BitBucket. To contribute, fork the project and send a pull request.

See the *Contributors* page for a list of people who have contributed to developing Monary.

Now install it!

On PyPI:

pypi.python.org/pypi/Monary

With pip:

```
$ pip install monary
```

The Future of Monary

Potential new features

❖ Release all our features

❖ Pandas integration

❖ Bulk updates, upserts

❖ Array extraction

❖ Logging

Contributing

Author: David J. C. Beach

License: Apache License Version 2.0

Open source on BitBucket:

bitbucket.org/djcbeach/monary

Thanks!

Our mentors at MongoDB:
- ❖ A. Jesse Jiryu Davis
- ❖ Jason Carey

The author of Monary:
- ❖ David J. C. Beach

Questions?

Source: bitbucket.org/djcbeach/monary

PyPI: pypi.python.org/pypi/Monary

Documentation: tinyurl.com/monary-docs

This Talk: github.com/Machyne/pygotham-monary