ASI DATA SCIENCE

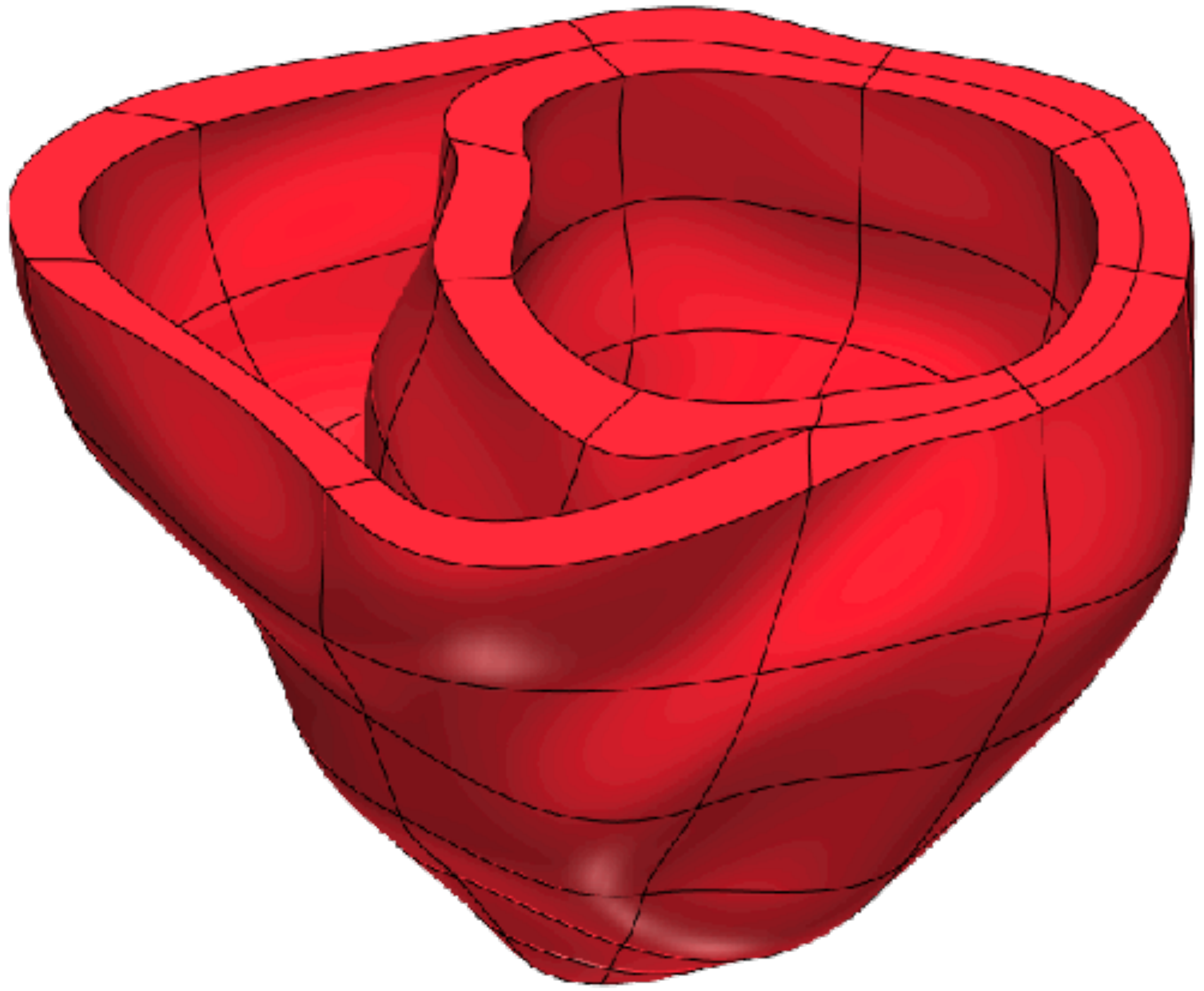
Using Spark from Python and Jupyter

Andrew Crozier
17th September 2018

QUEEN'S UNIVERSITY BELFAST

EST.º 1845

KING'S College LONDON



SALUS ÆGROTI SUPREMA LEX • MEDIZINISCHE UNIVERSITÆT GRAZ •

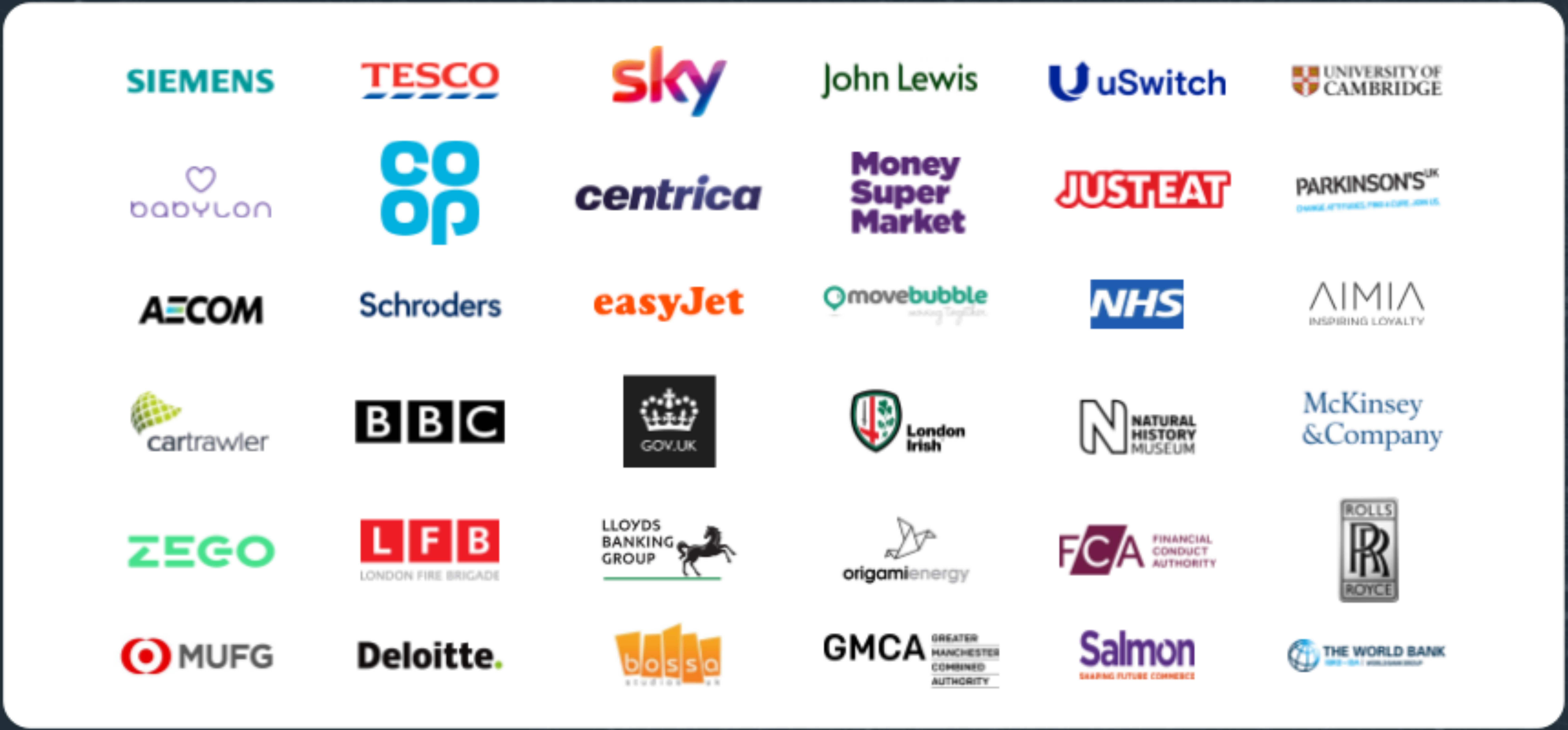
ASi

acroz

acroz@asidatascience.com

ASI helps apply Artificial Intelligence to solve business and policy problems.

We work with organisations in every sector



ASI

SHERLOCKML

Define

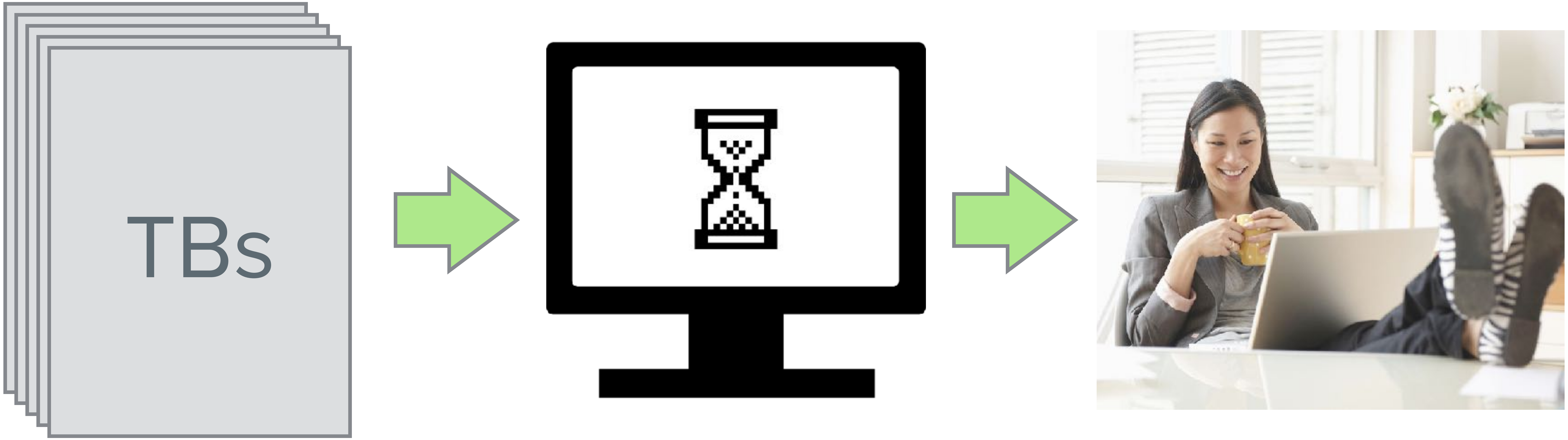
Upload, clean and find patterns in large datasets

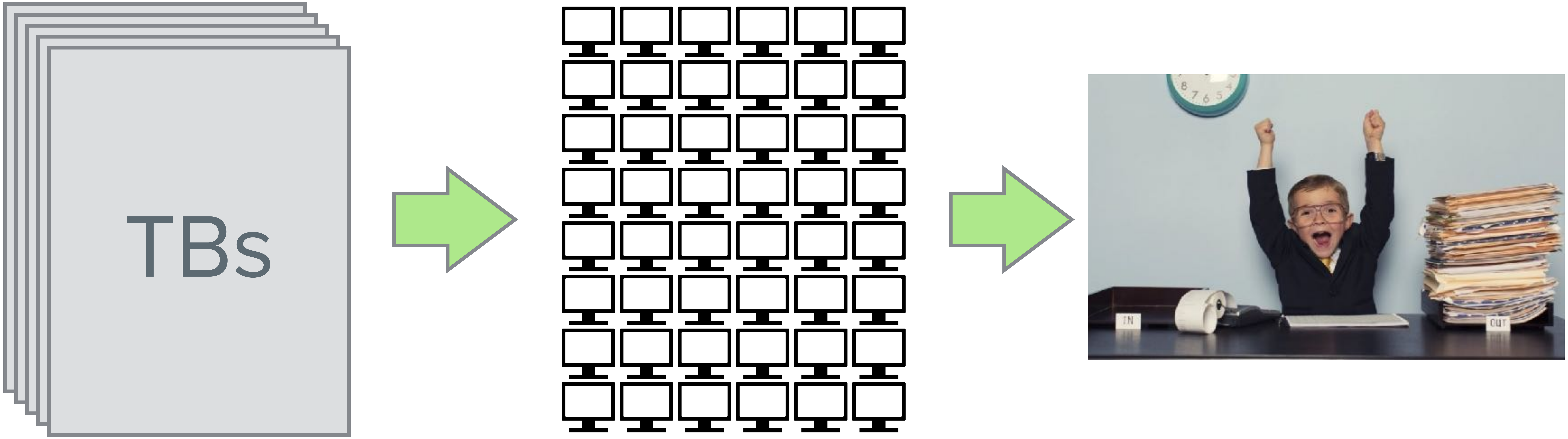
Develop

Design and test AI models on your data.

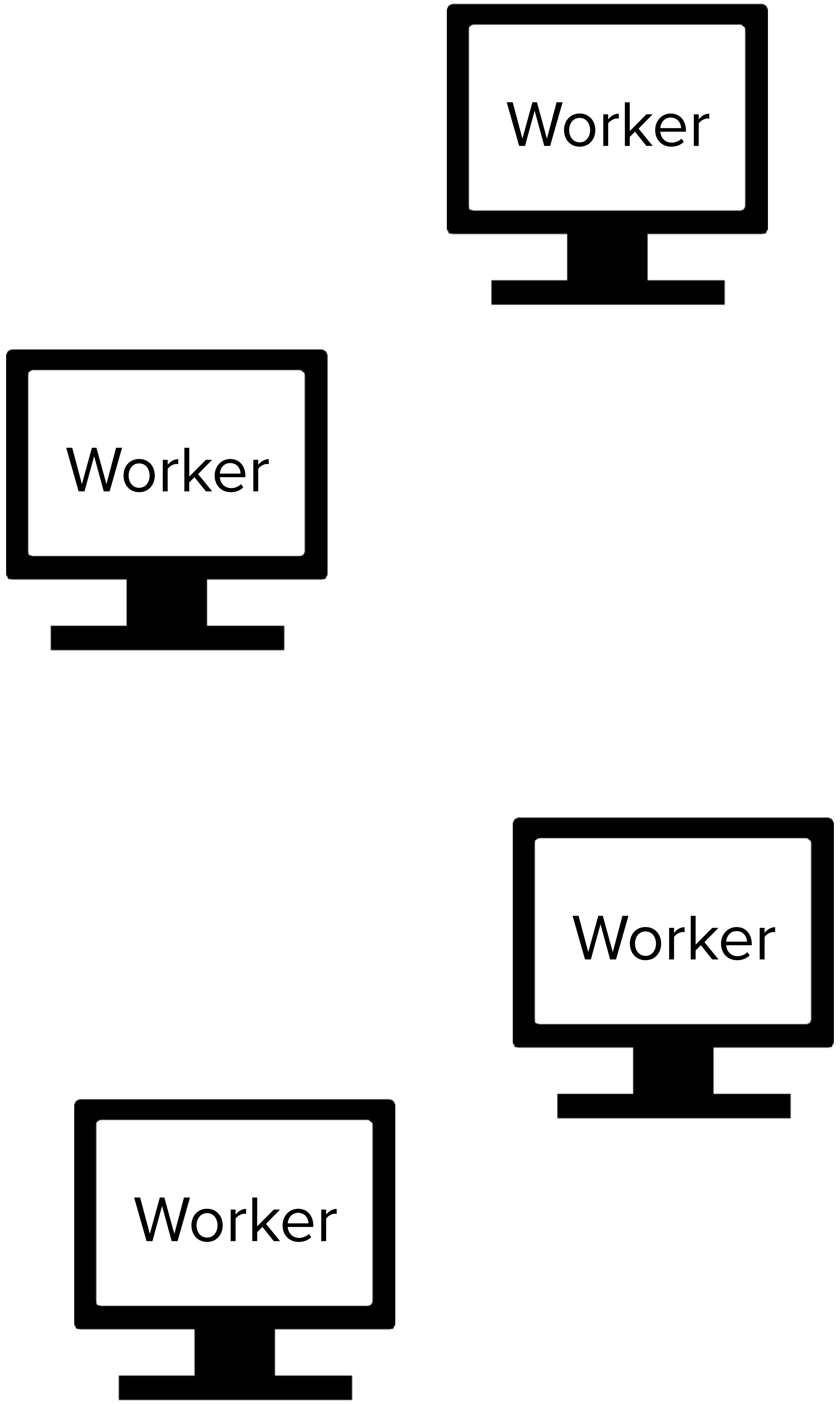
Deploy

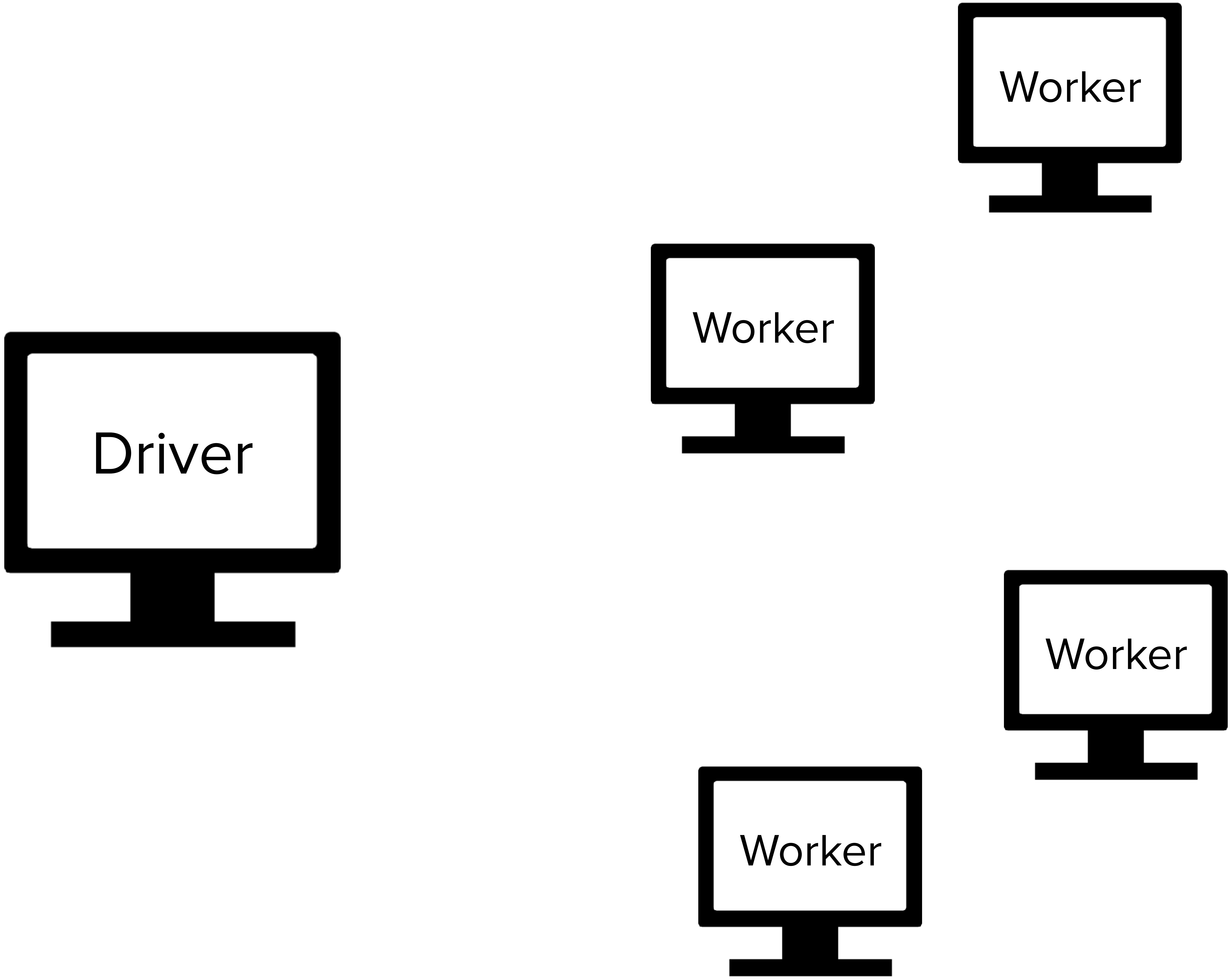
Generate a report or deploy through an API


TBs


TBs
IN
OUT

APACHE
Spark
TM


Worker
Worker
Worker
Worker



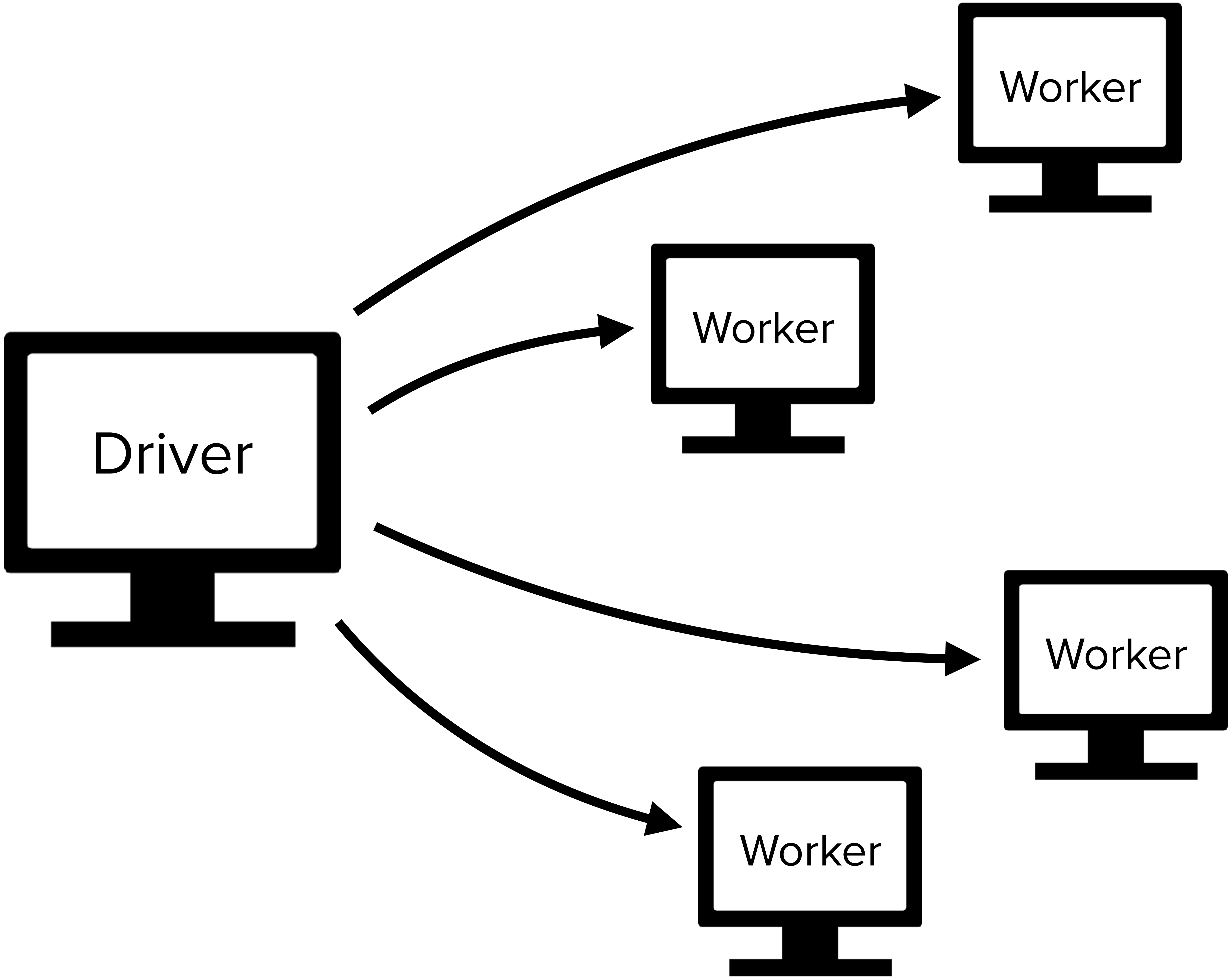
Driver

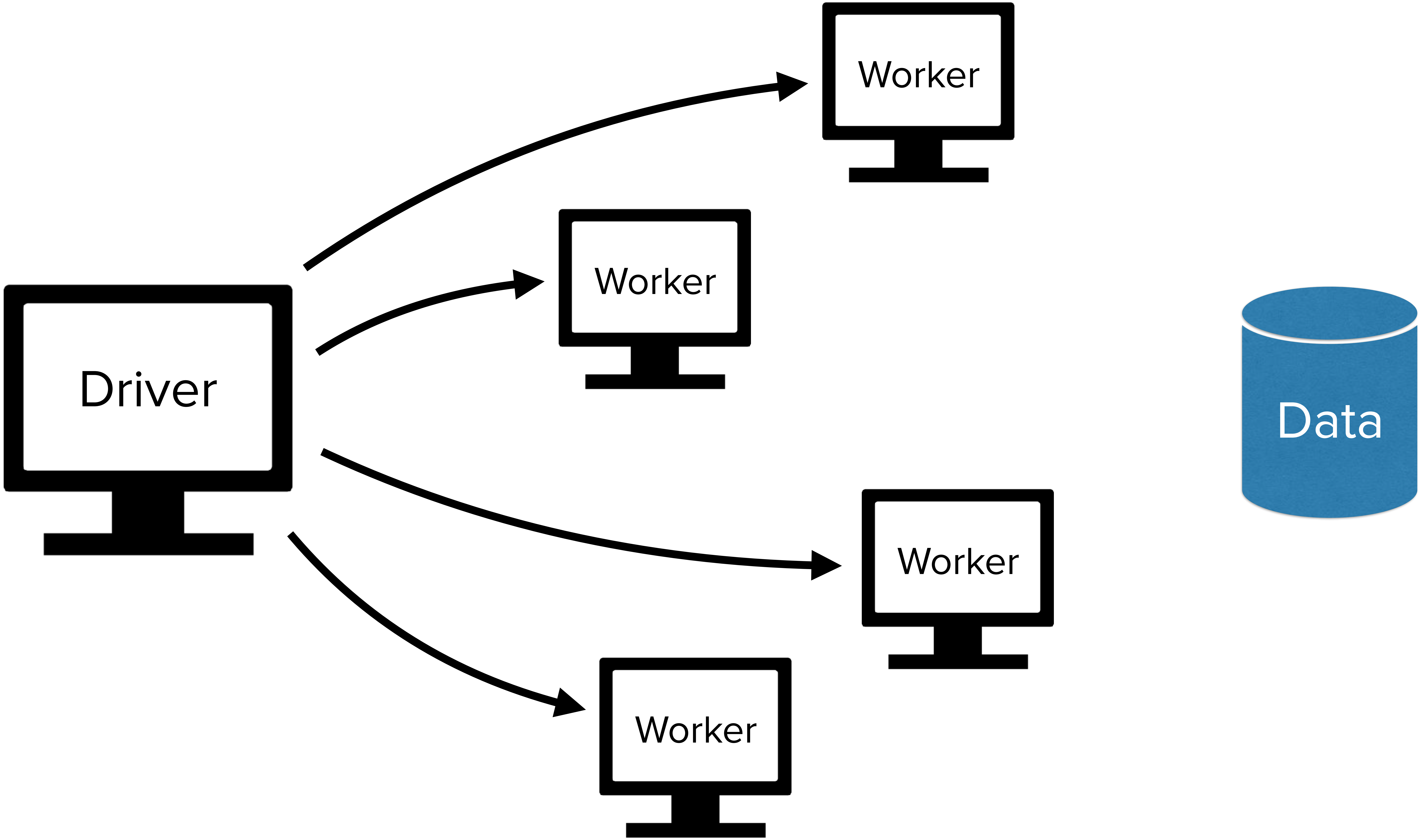
Worker

Worker

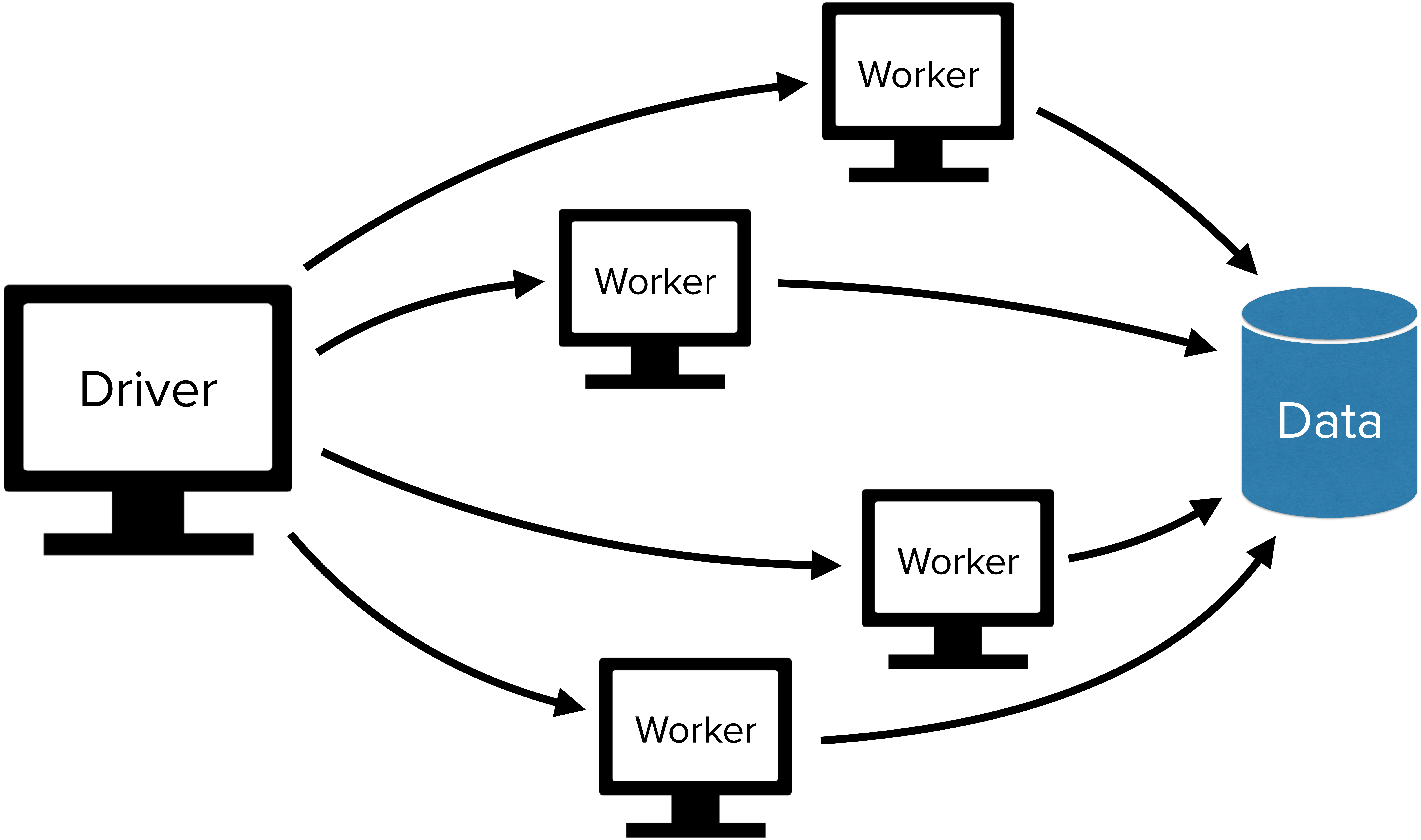
Worker

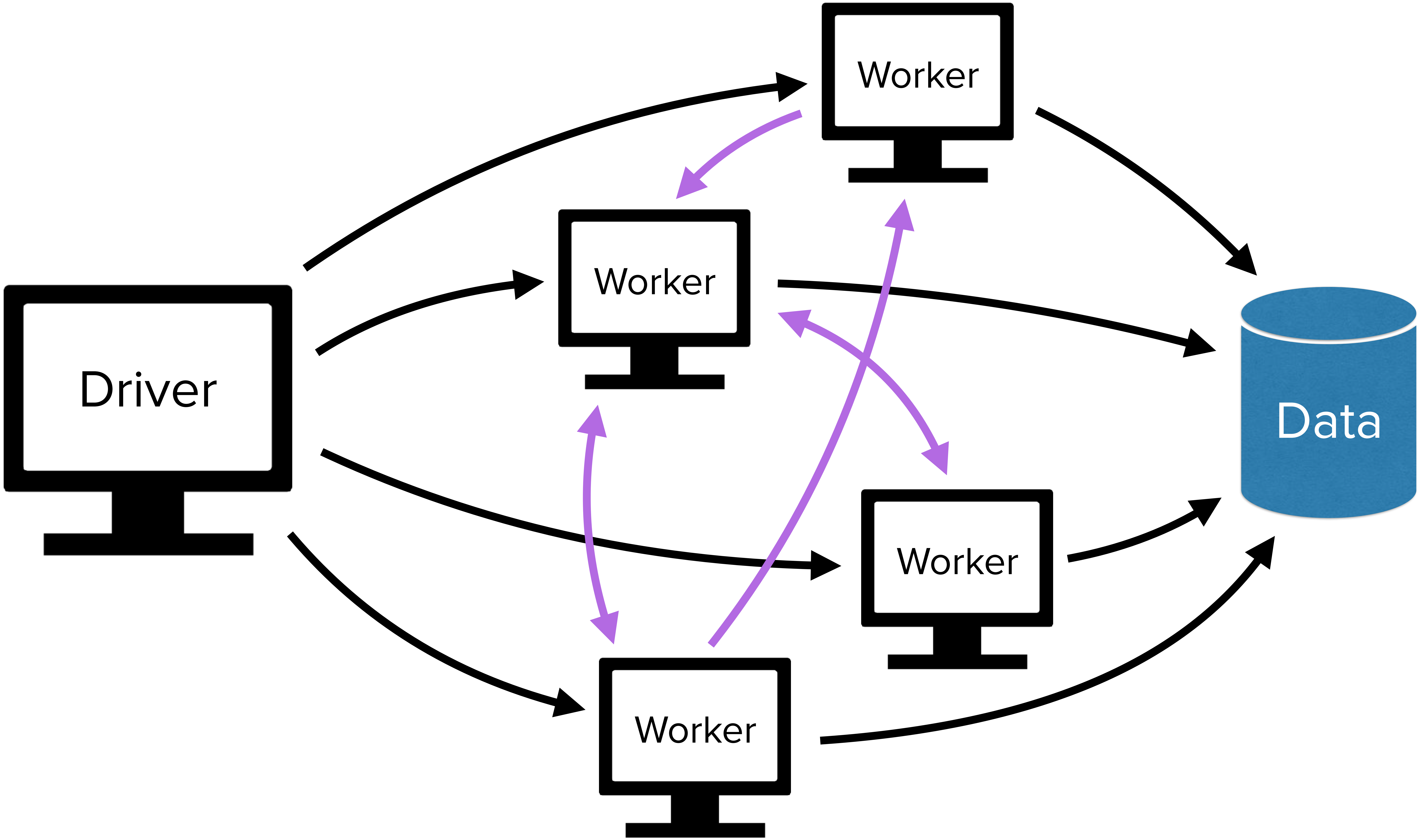
Worker

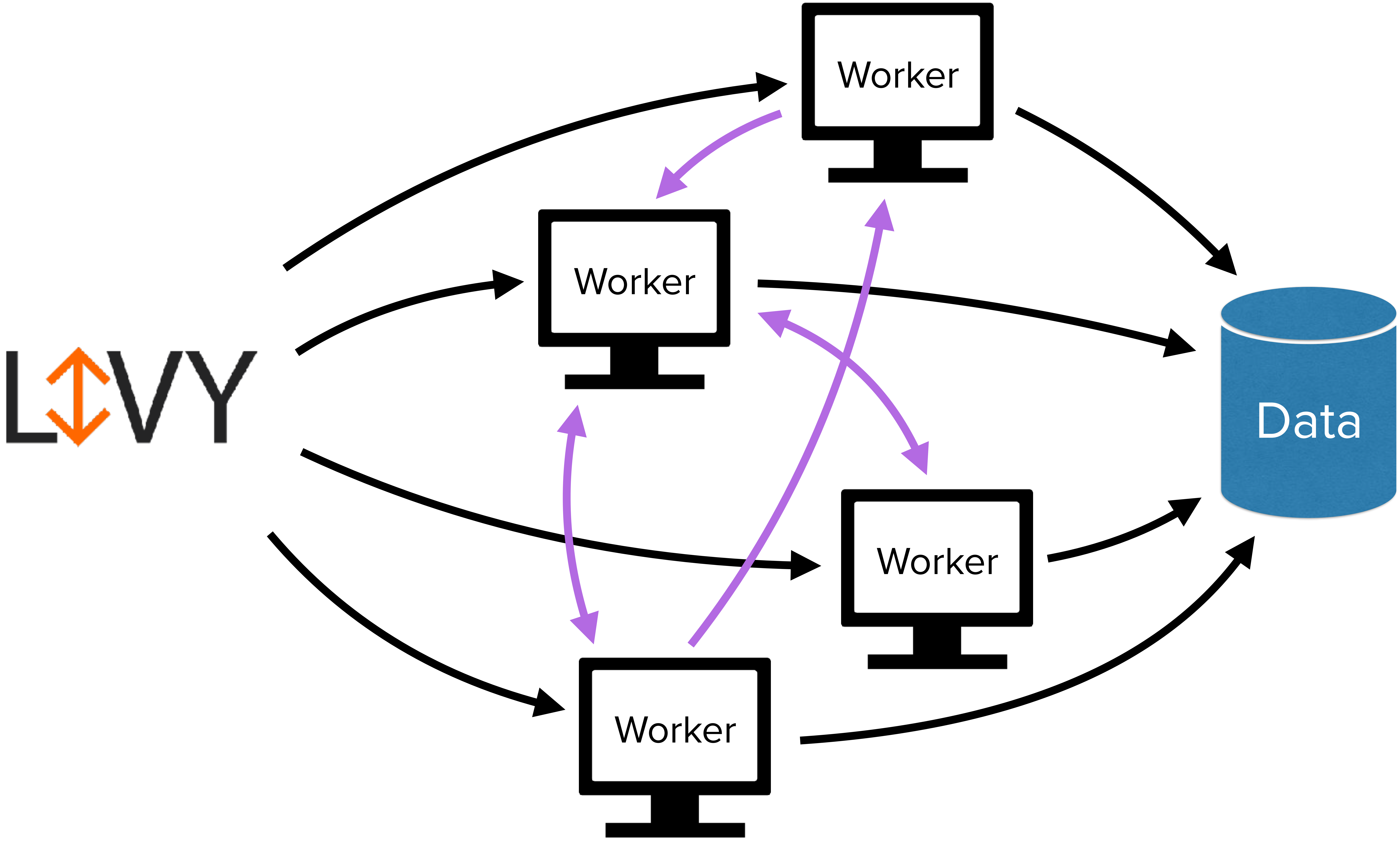

Worker
Worker
Driver
Worker
Worker

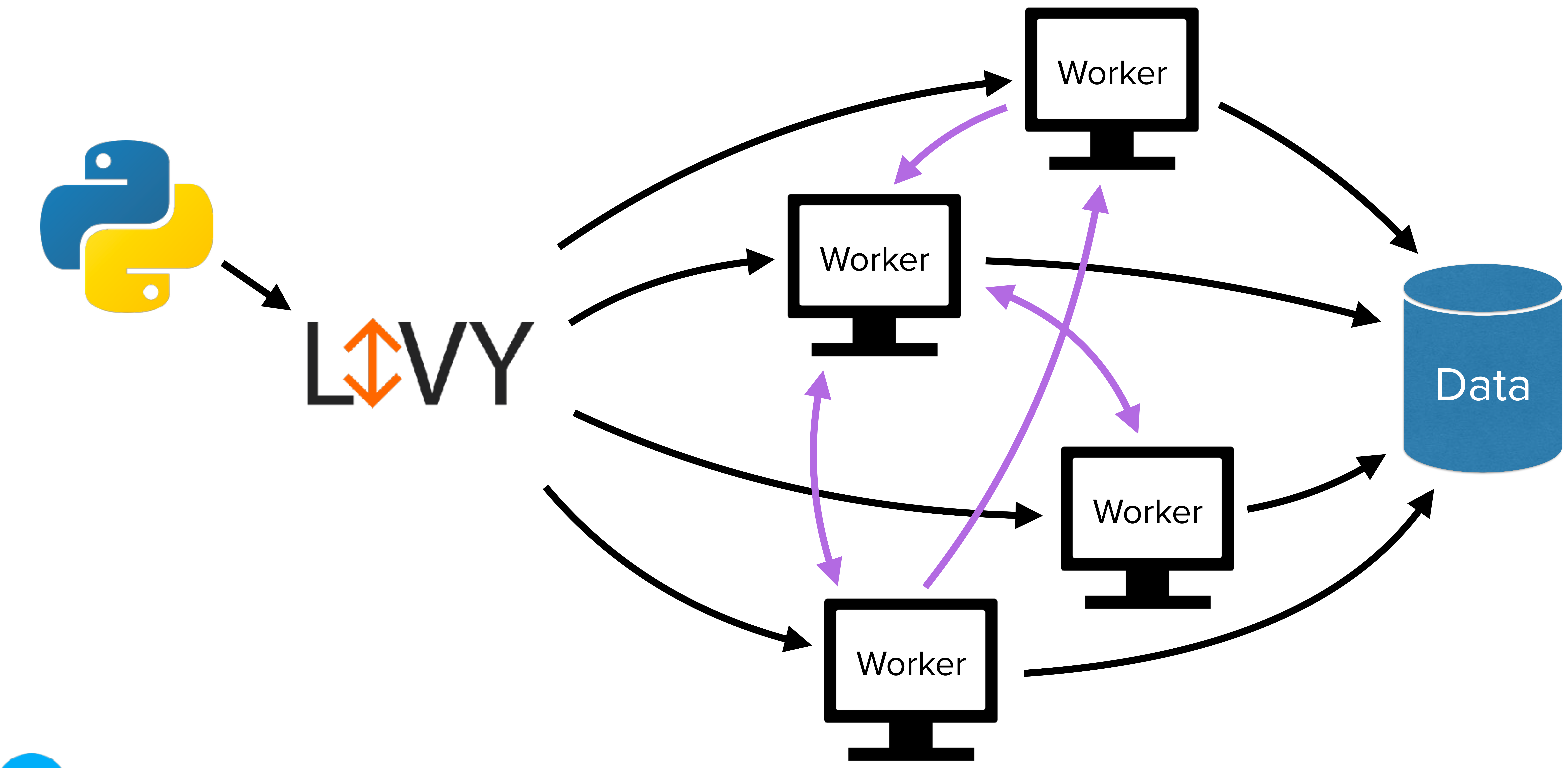


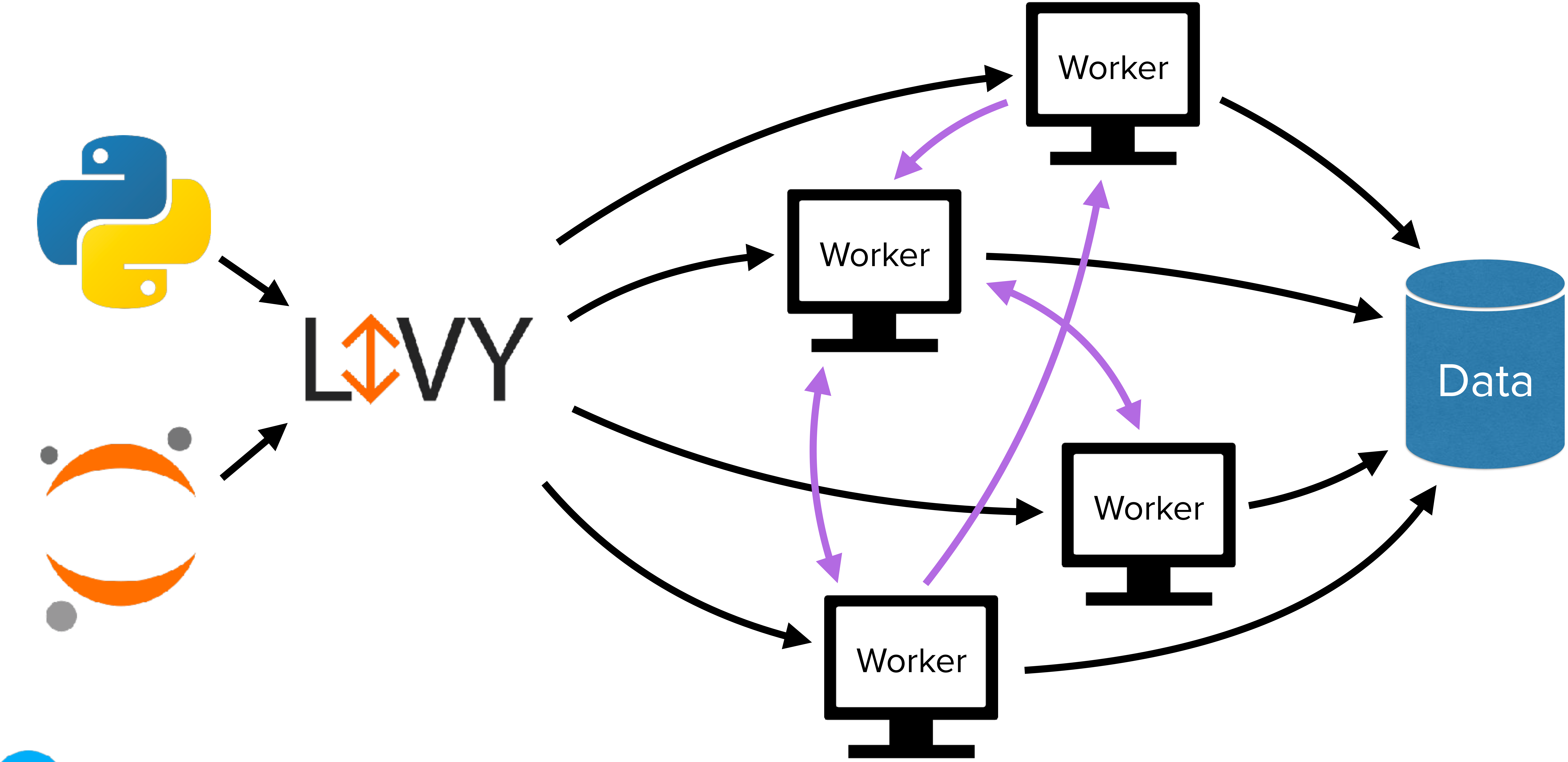
Worker
Worker
Driver
Data
Worker
Worker


Worker
Worker
Driver
Data
Worker
Worker


Worker
Worker
Driver
Data
Worker
Worker

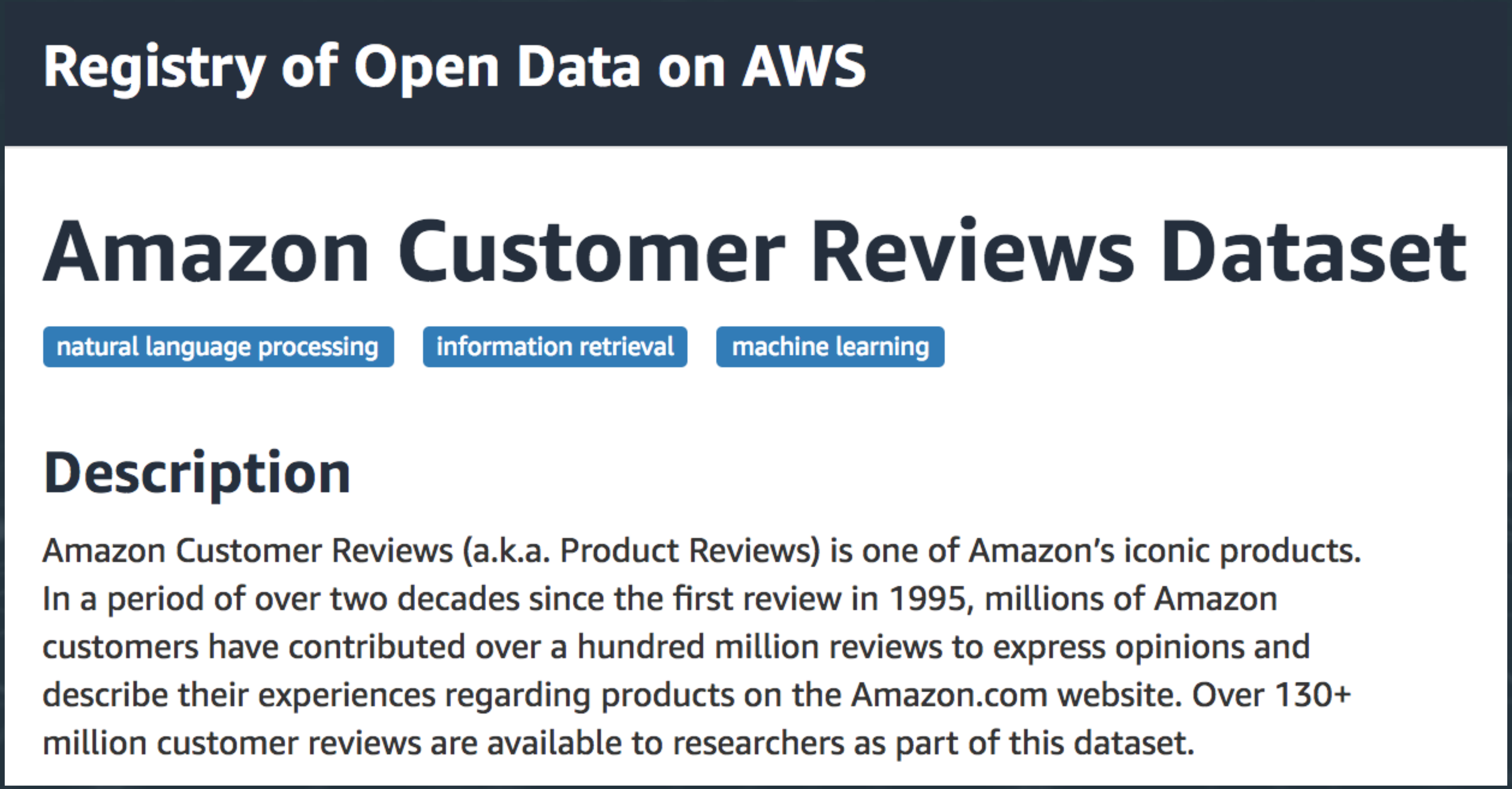

LIVY
Worker
Worker
Data
Worker
Worker


LIVY
Worker
Worker
Worker
Worker
Data


LIVY
Worker
Worker
Worker
Worker
Data

Synopsis

- Connect to a Spark cluster from Jupyter
- Load data into Spark
- Basics of data manipulation in Spark
- Getting data out of Spark
- Introduction to Spark machine learning tools
- Using Spark from Python scripts and web apps



Registry of Open Data on AWS

Amazon Customer Reviews Dataset

natural language processing | information retrieval | machine learning

Description

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. Over 130+ million customer reviews are available to researchers as part of this dataset.

https://registry.opendata.aws/amazon-reviews/

amazon
web services

sparkmagic

```
pip install sparkmagic
```

jupyter-incubator/sparkmagic

Demo

pylivy

```
pip install livy
```

acroz/pvlivy

```python
from livy import LivySession
from livy import LivySession


with LivySession("http://localhost:8998") as session:
```

```python
from livy import LivySession


with LivySession("http://localhost:8998") as session:

    session.run(
        "df = spark.read.parquet('s3://amazon-reviews-pds/parquet/')"
    )
```

```python
from livy import LivySession


with LivySession("http://localhost:8998") as session:

    session.run(
        "df = spark.read.parquet('s3://amazon-reviews-pds/parquet/')"
    )

    # Any output from the remote interpreter is displayed
    session.run("df.printSchema()")
```

```python
from livy import LivySession
from textwrap import dedent

with LivySession("http://localhost:8998") as session:

    session.run(
        "df = spark.read.parquet('s3://amazon-reviews-pds/parquet/')"
    )

    # Any output from the remote interpreter is displayed
    session.run("df.printSchema()")

    # Multi-line snippets can be passed
    session.run(dedent("""
        star_rating_counts = df.groupBy('star_rating').count()
        ordered_counts = star_rating_counts.orderBy('star_rating')
    """))
```

```python
from livy import LivySession
from textwrap import dedent

with LivySession("http://localhost:8998") as session:

    session.run(
        "df = spark.read.parquet('s3://amazon-reviews-pds/parquet/')"
    )

    # Any output from the remote interpreter is displayed
    session.run("df.printSchema()")

    # Multi-line snippets can be passed
    session.run(dedent("""
        star_rating_counts = df.groupBy('star_rating').count()
        ordered_counts = star_rating_counts.orderBy('star_rating')
    """))

    counts = session.read("ordered_counts")
```

Dashboards!

Dash
by plotly

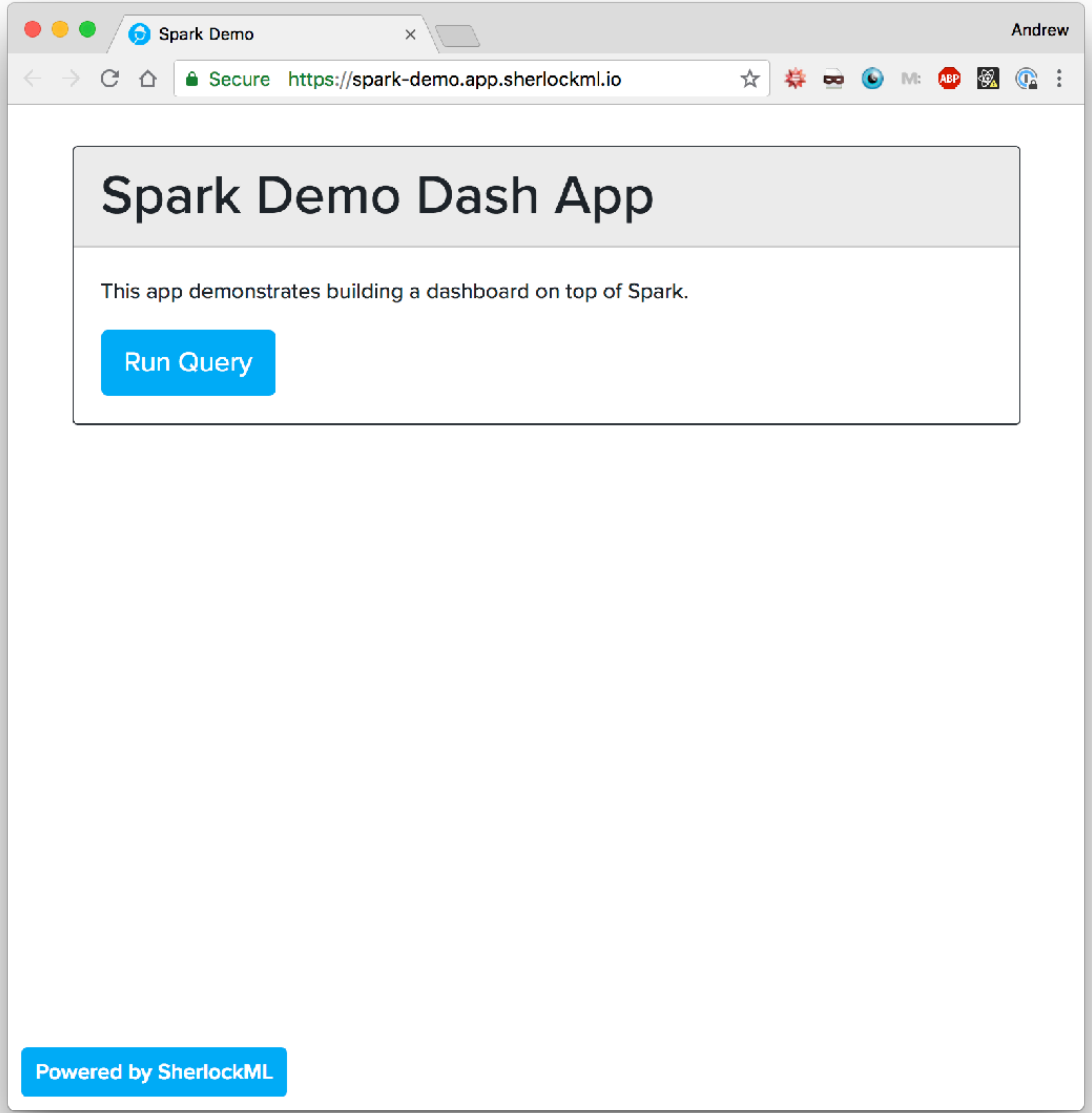
+

RQ

+

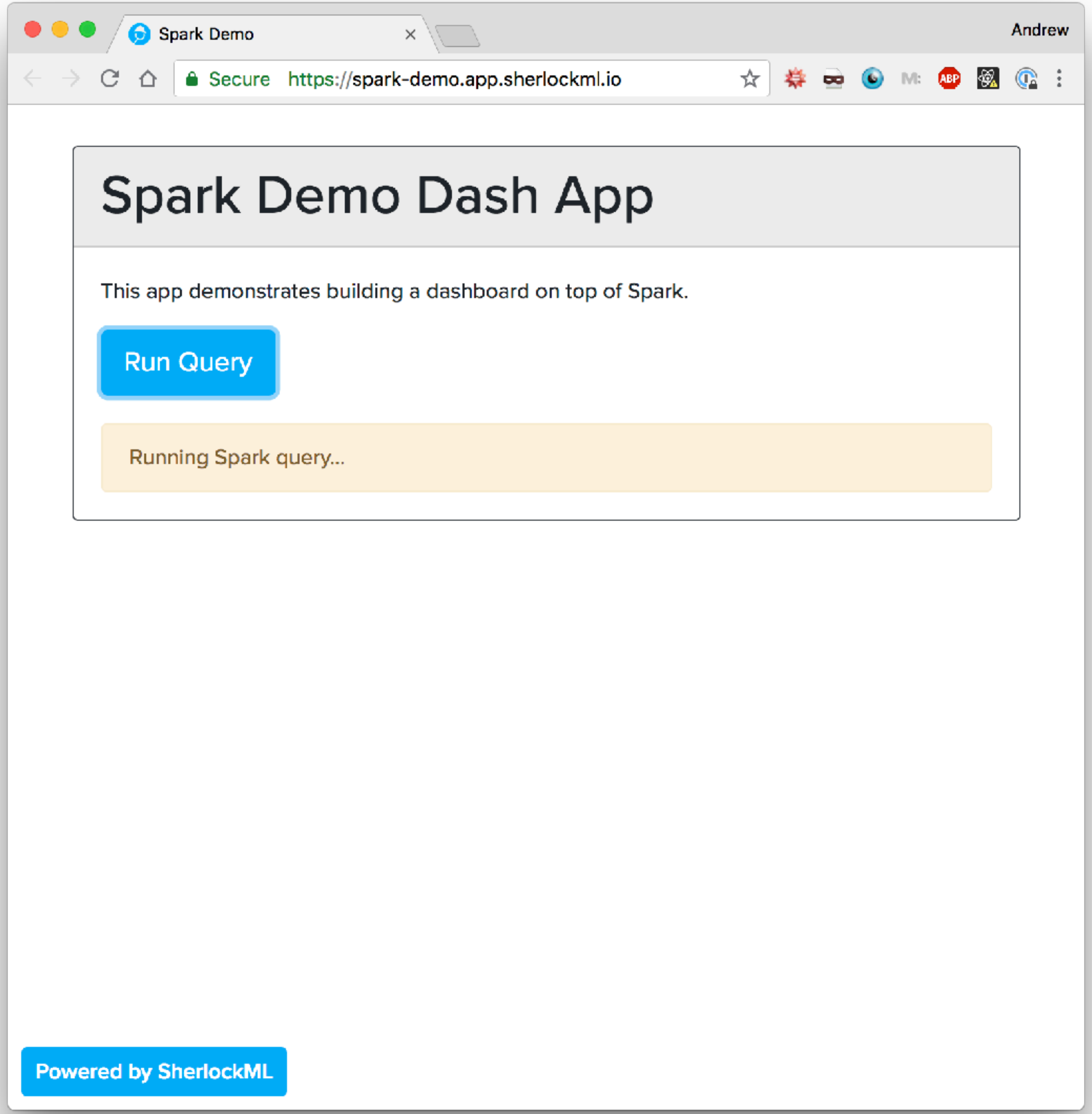
LIVY

Demo



Spark Demo Dash App

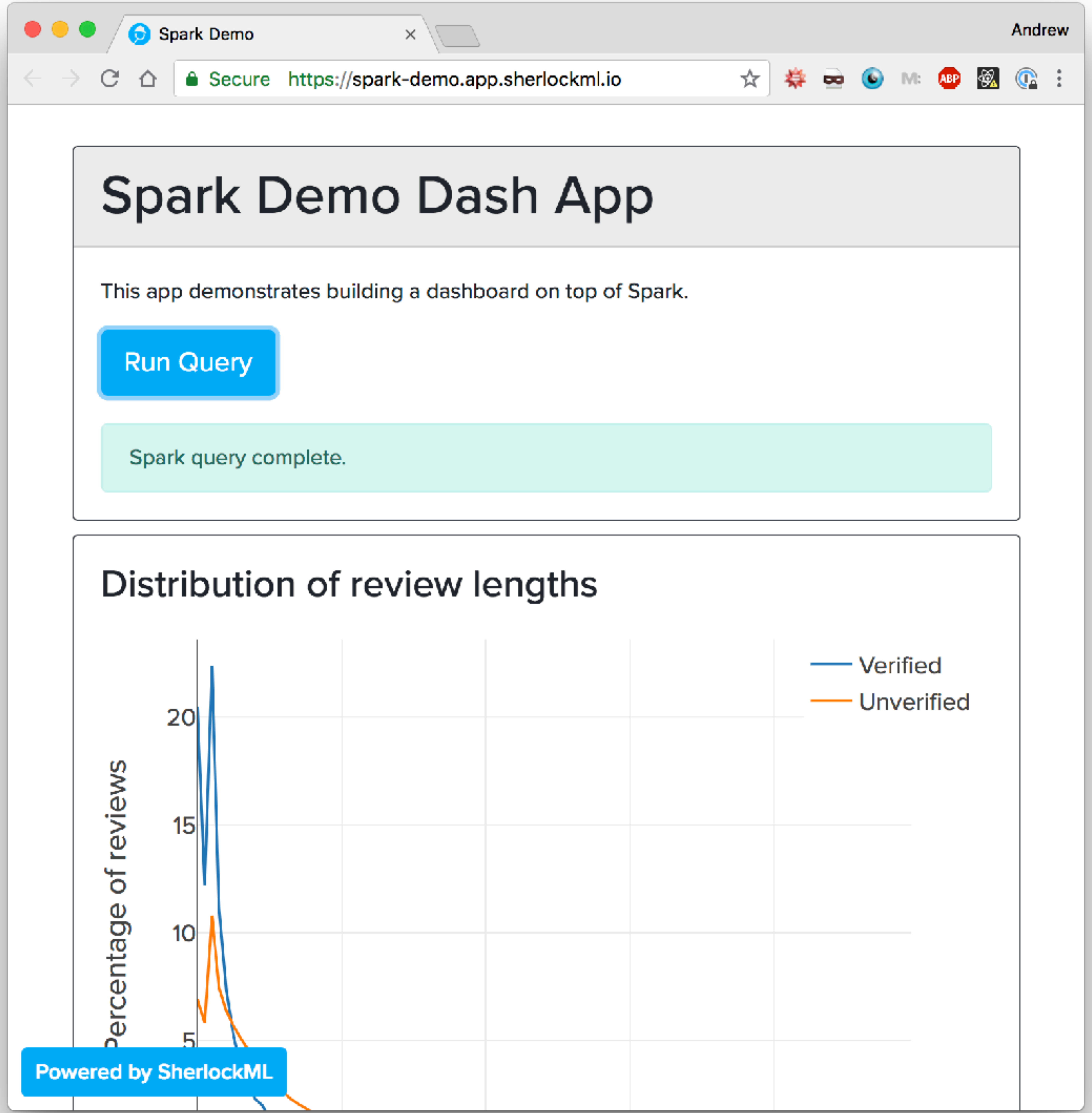
This app demonstrates building a dashboard on top of Spark.



Spark Demo Dash App

This app demonstrates building a dashboard on top of Spark.

Running Spark query...

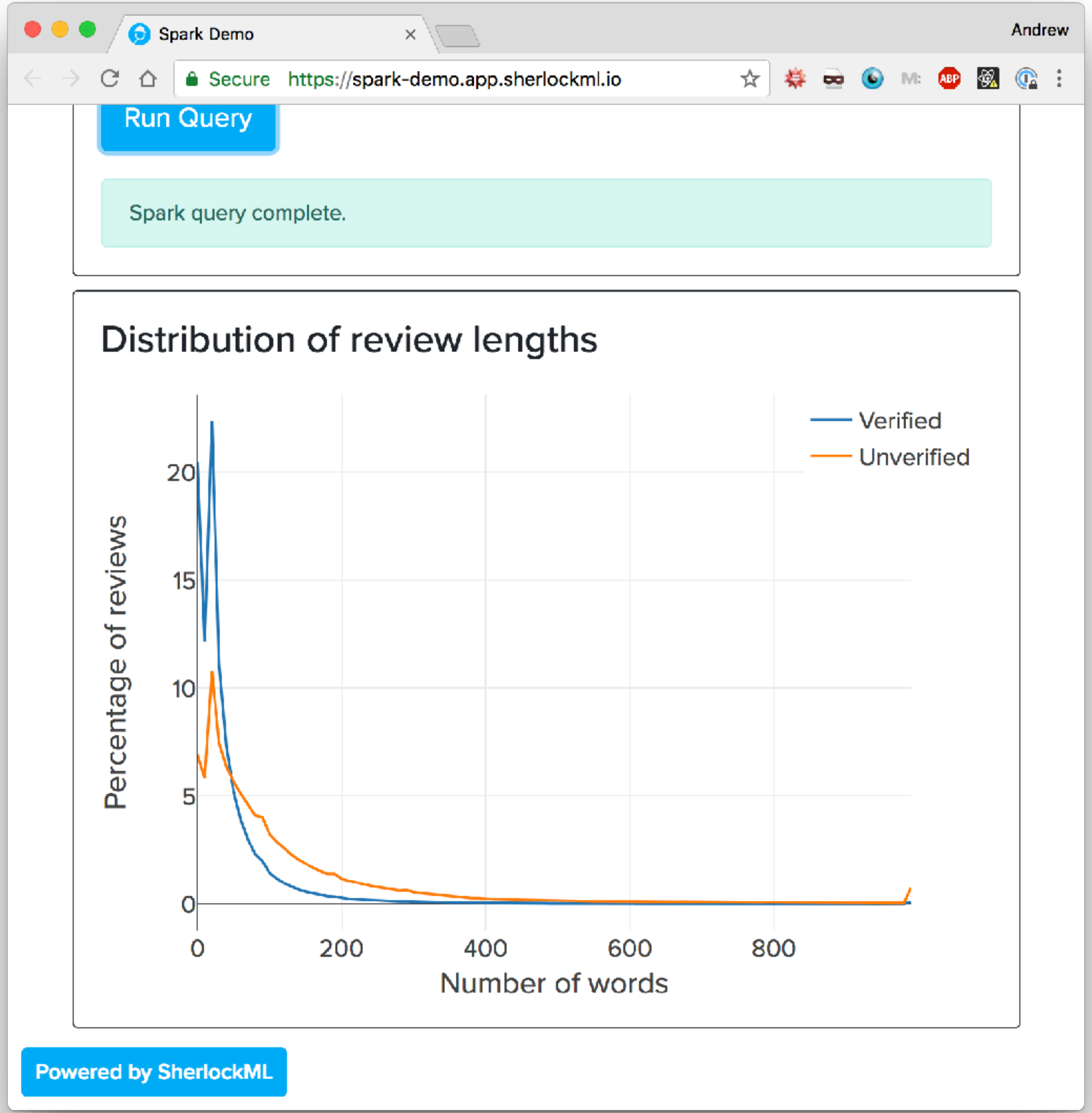


Spark Demo Dash App

This app demonstrates building a dashboard on top of Spark.

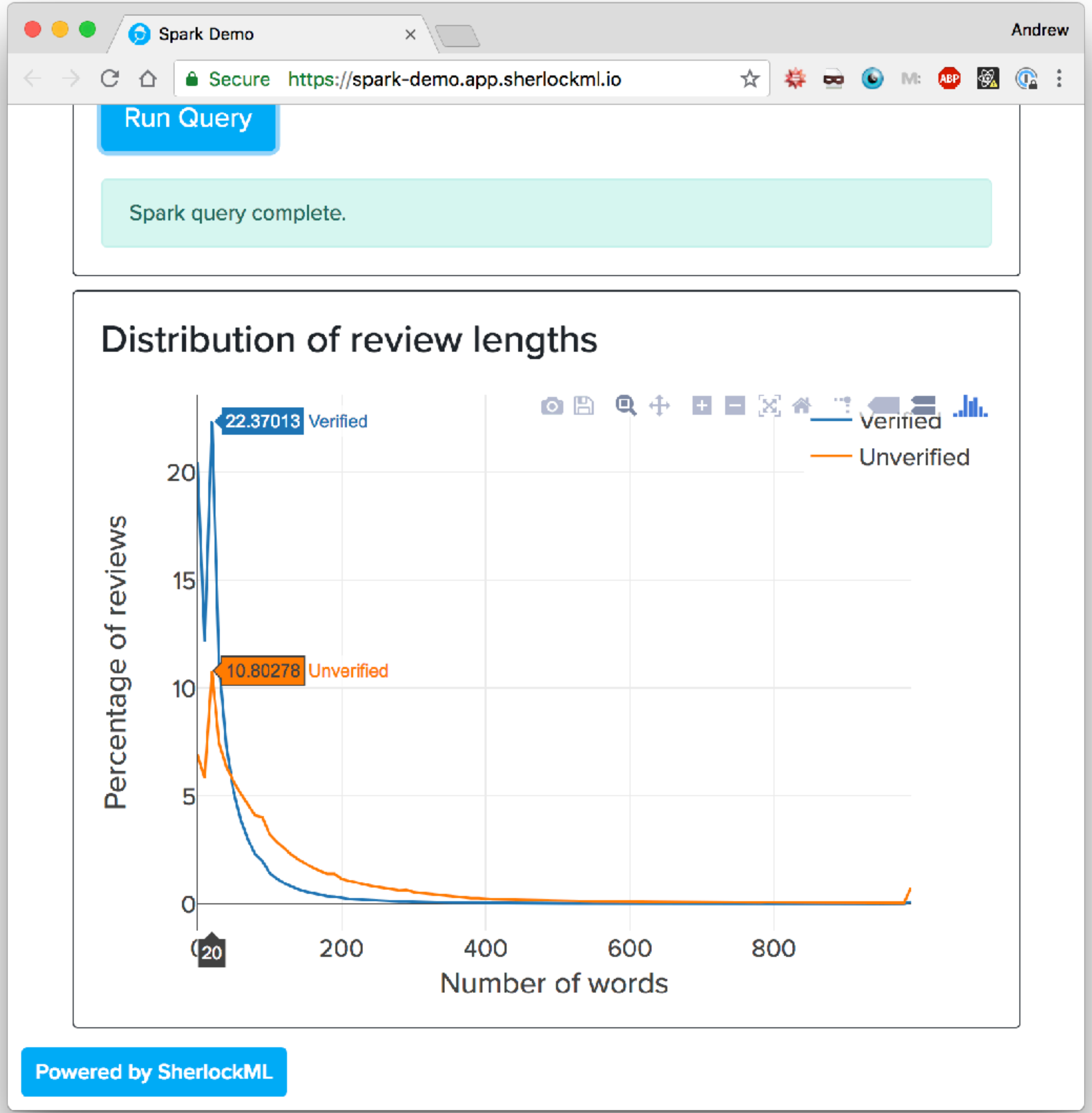
Spark query complete.

Distribution of review lengths



Spark query complete.

Distribution of review lengths



Spark query complete.

Distribution of review lengths

https://spark.apache.org/docs/latest/

https://spark.apache.org/docs/latest/ml-guide.html

jupyter-incubator/sparkmagic

acroz/pvlivy

https://sherlockml.com

acroz

acroz@asidatascience.com