Dask Internals and Dashboard

Paweł Kordek
SOFTWARE ENGINEER

@pawel_kordek   https://kordek.github.io

Internals

Representation

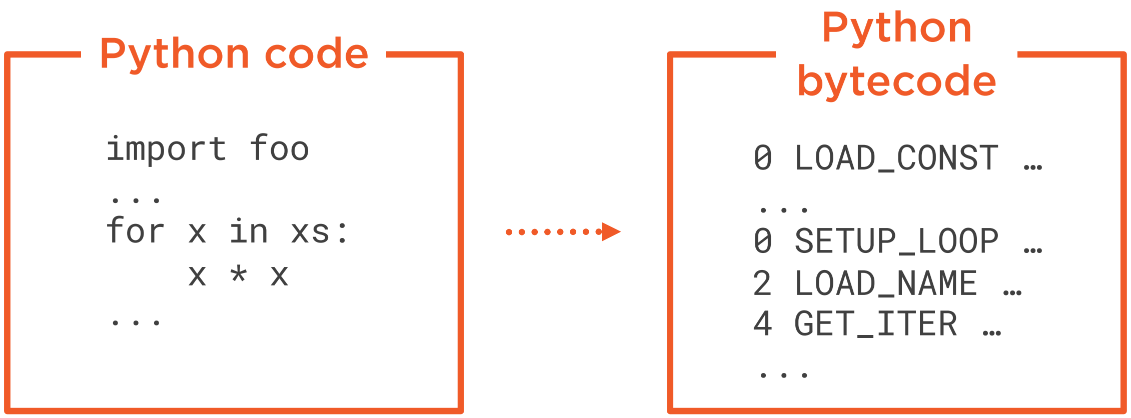
Runtime

Visualization

Big Picture

Python code

```
import foo
...
for x in xs:
    x * x
...
```

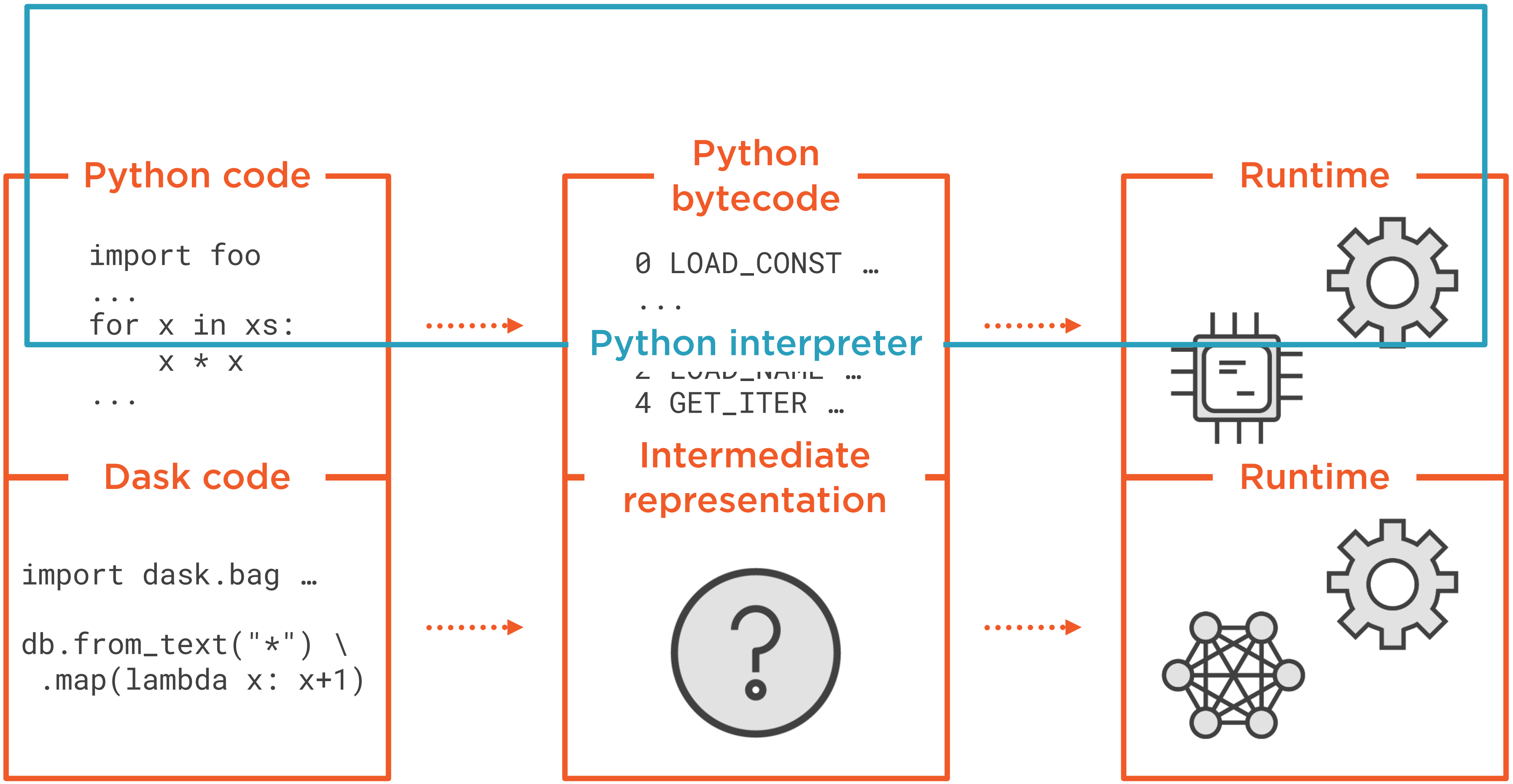


Python code

```
import foo
...
for x in xs:
    x * x
...
```

Python bytecode

```
0 LOAD_CONST …
...
0 SETUP_LOOP …
2 LOAD_NAME …
4 GET_ITER …
...
```



Python code

```
import foo
...
for x in xs:
    x * x
...
```

Python bytecode

```
0 LOAD_CONST …
...
2 LOAD_NAME …
4 GET_ITER …
```

Python interpreter

Runtime

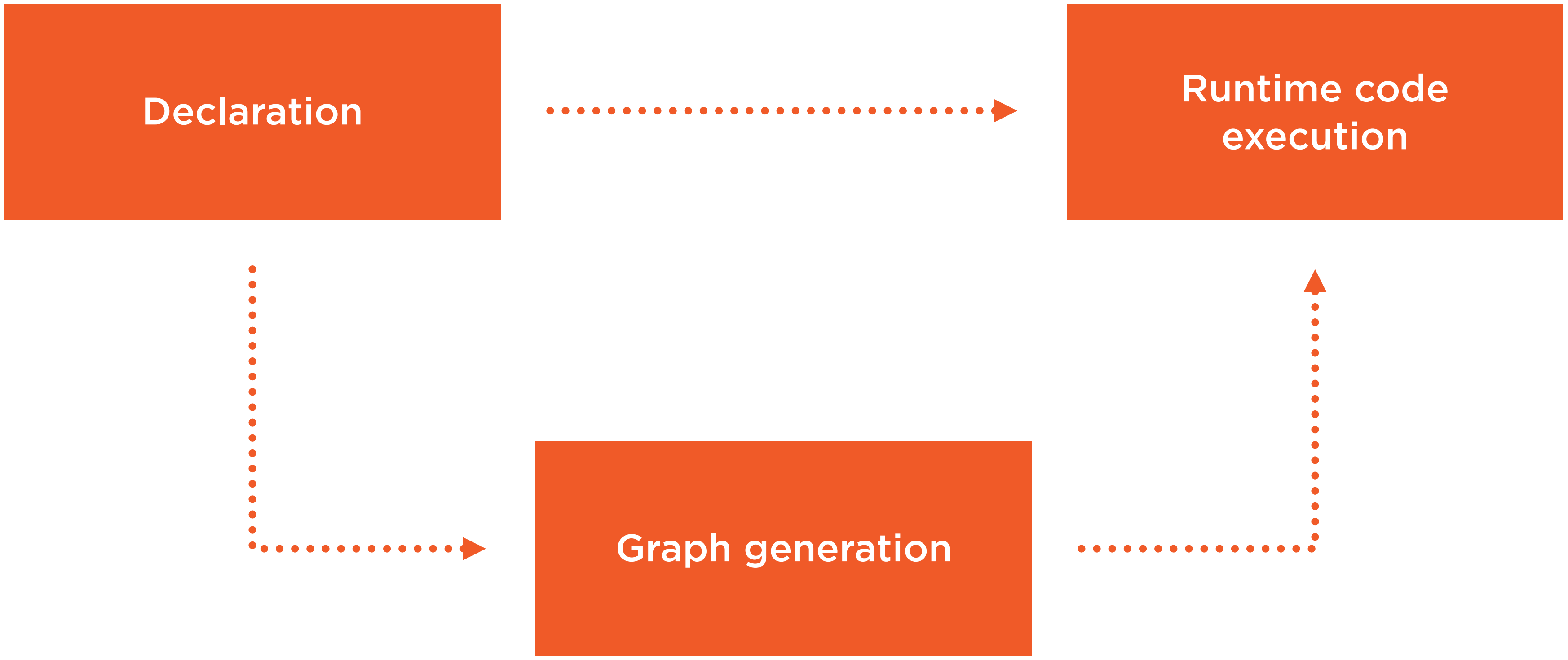
Dask code

```
import dask.bag …

db.from_text("*") \
 .map(lambda x: x+1)
```

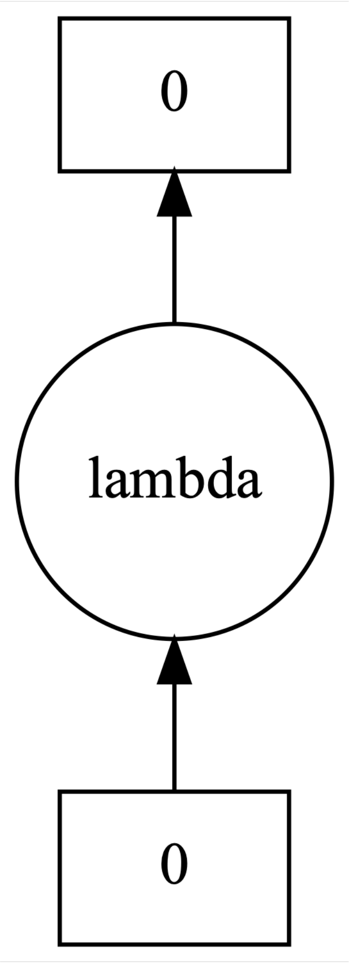
Intermediate representation

Runtime

Representation


Declaration
Runtime code execution
Graph generation

```python
import dask.bag as db

seq = [1, 2, 3, 4]
bag = db.from_sequence(
          seq, npartitions=1
      )

bag.map(lambda x: x + 1)
```

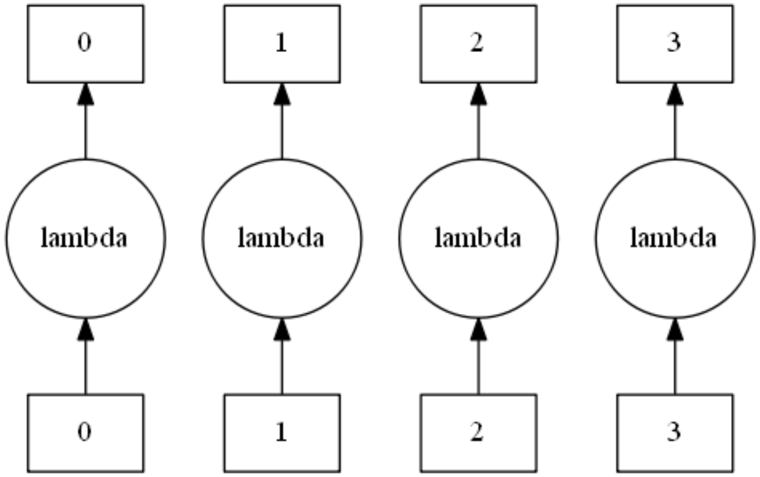


Partitions

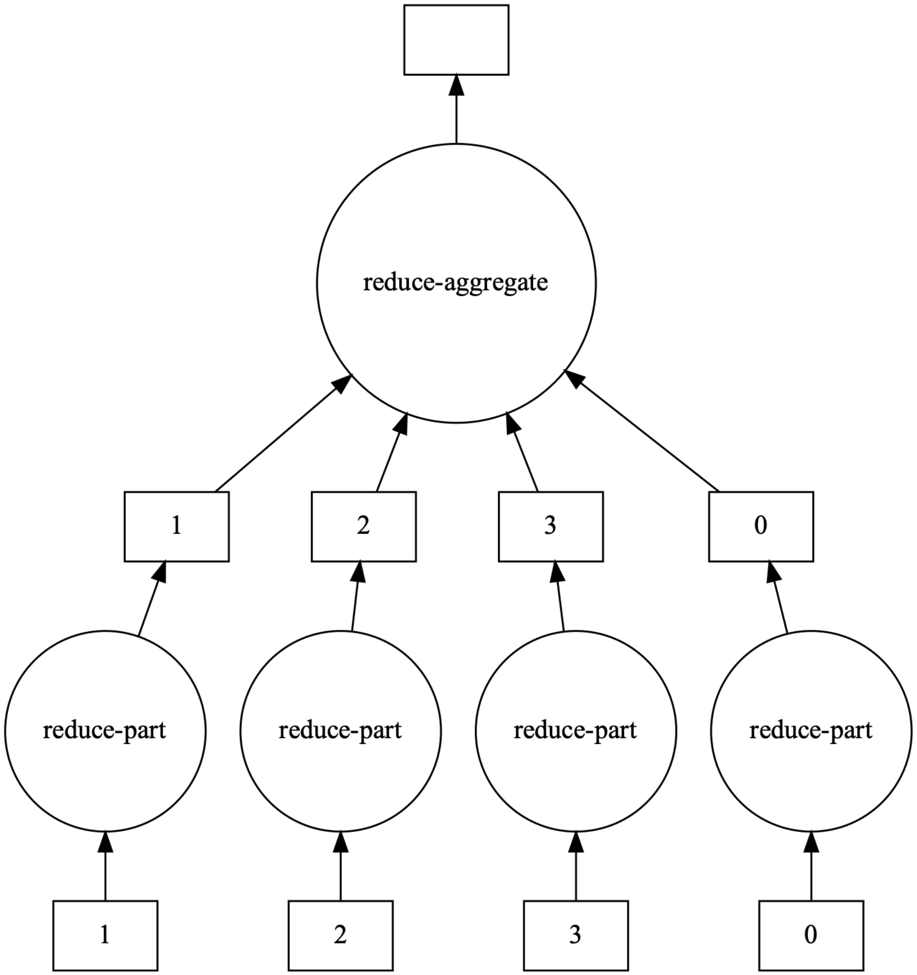
Logical portions of data

Affect parallelism of the application

```python
import dask.bag as db

seq = [1, 2, 3, 4]
bag = db.from_sequence(
          seq, npartitions=4
      )

bag.map(lambda x: x + 1)
```



```python
import dask.bag as db

seq = [1, 2, 3, 4]
bag = db.from_sequence(
          seq, npartitions=4
      )

bag.fold(lambda x, y: x + y)
```

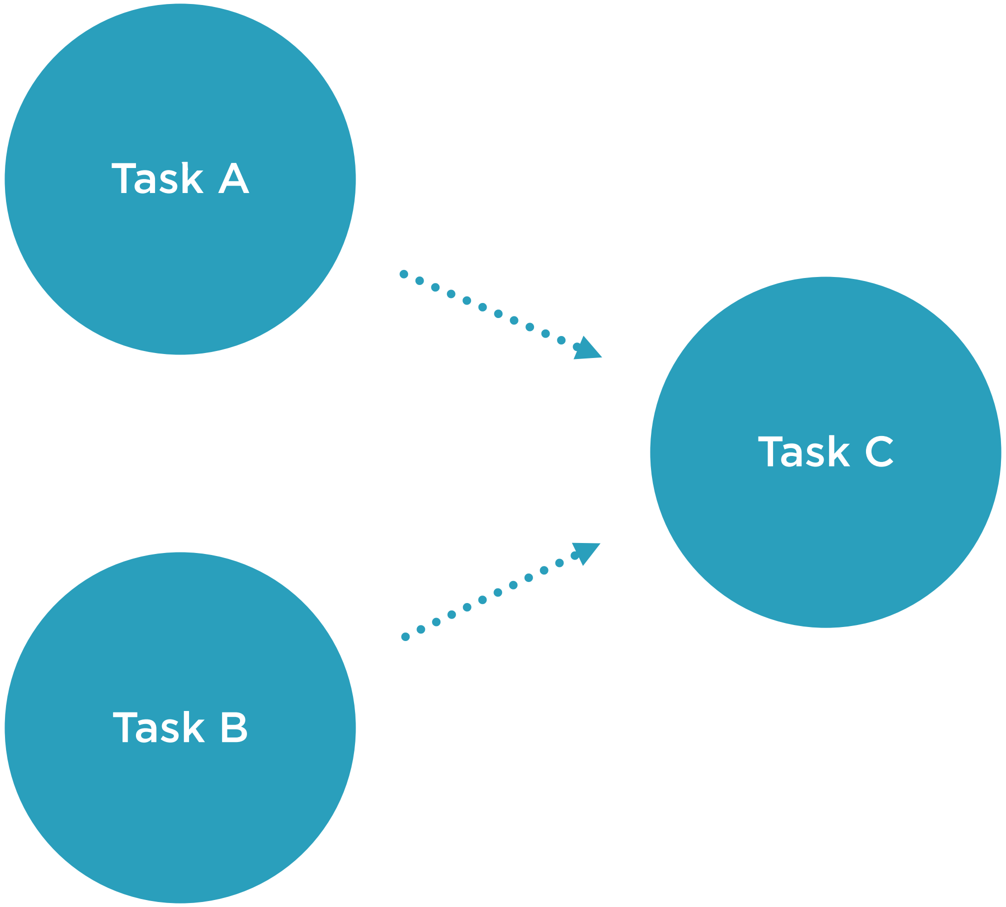


Graphs – Part of the Core

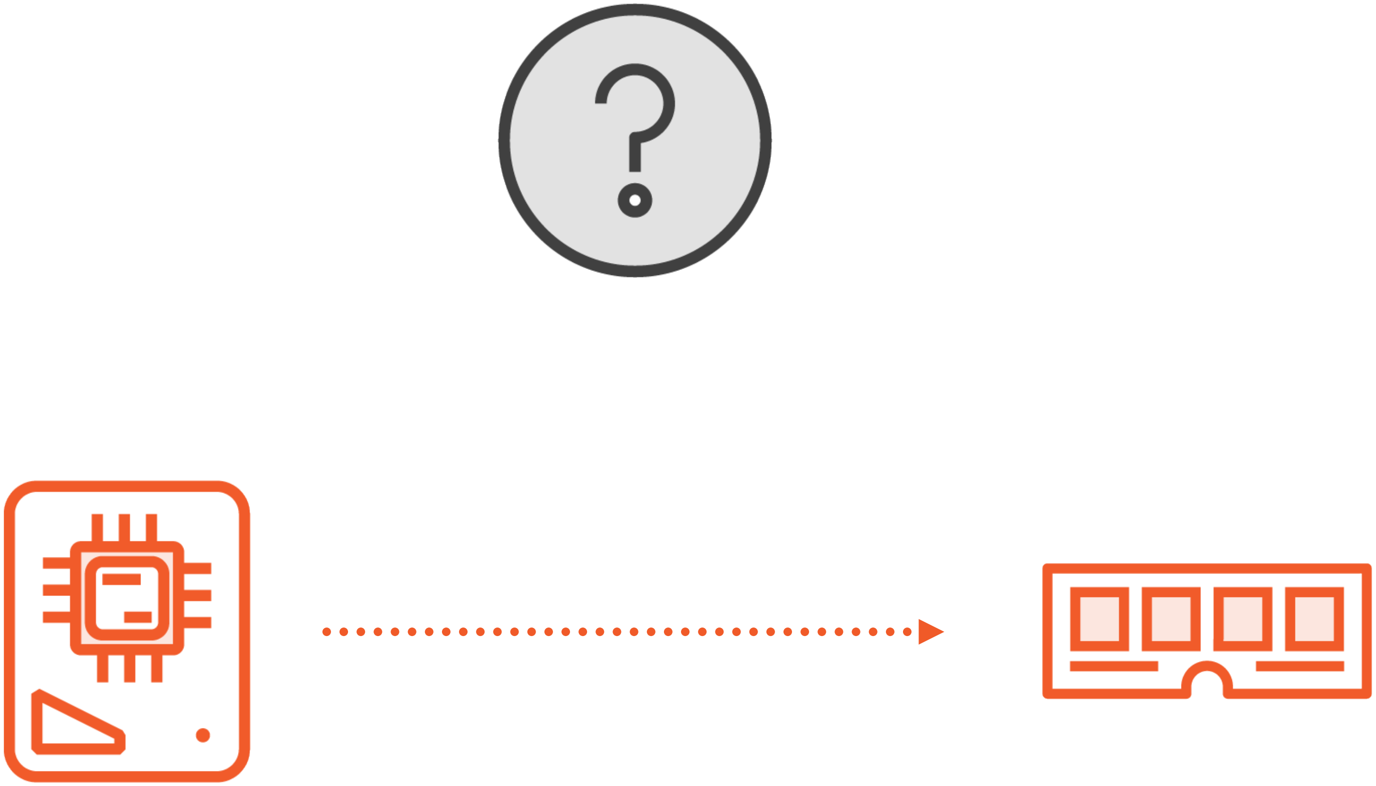
Directed

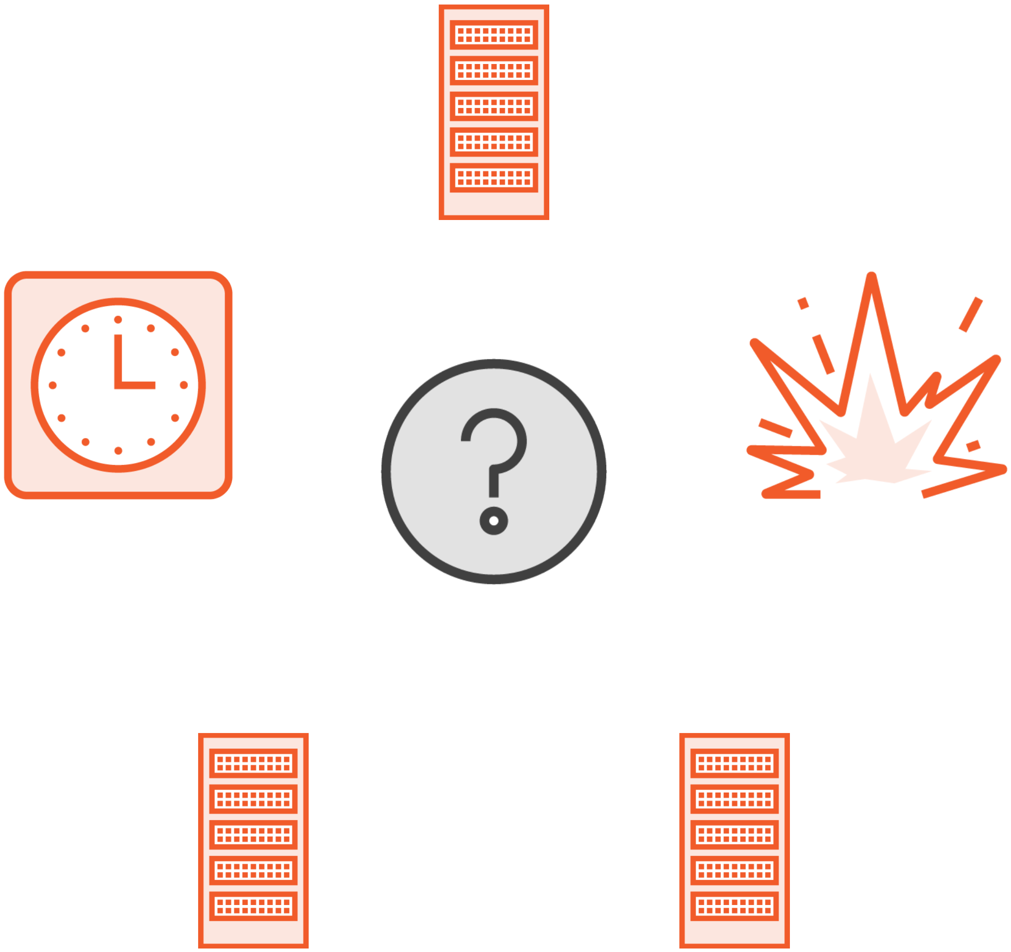
No cycles allowed

Directed Acyclic Graph

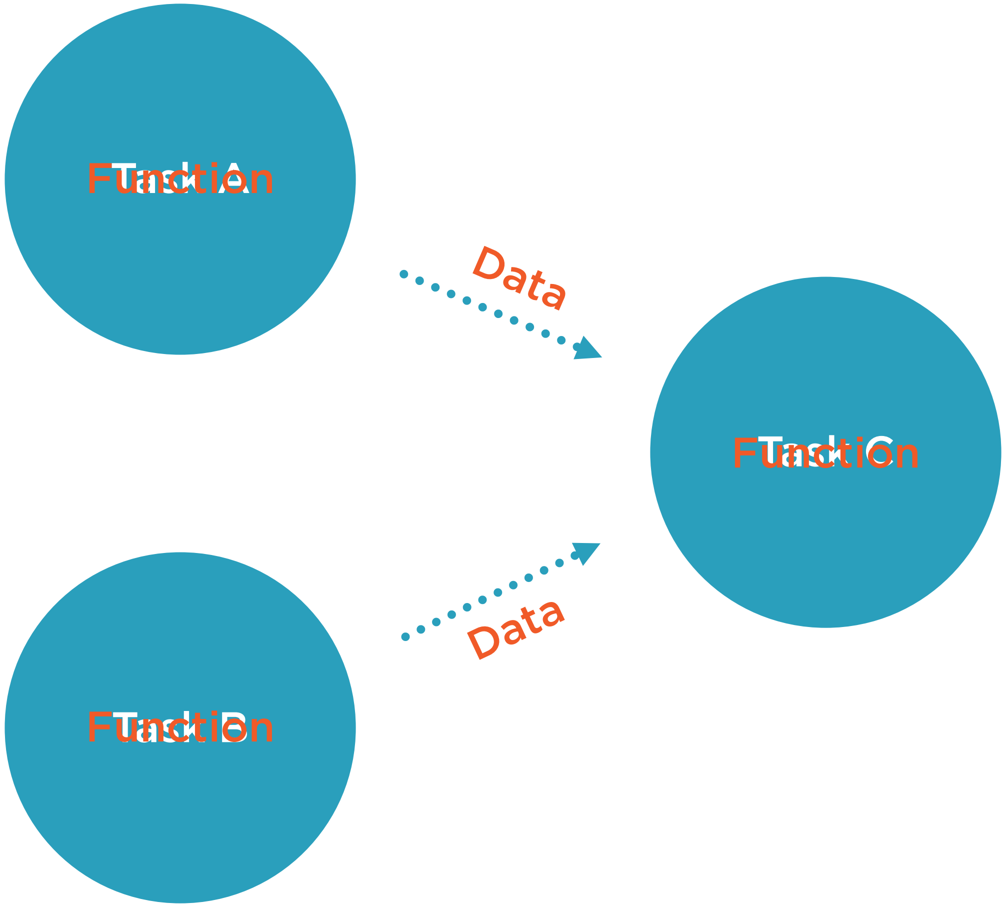


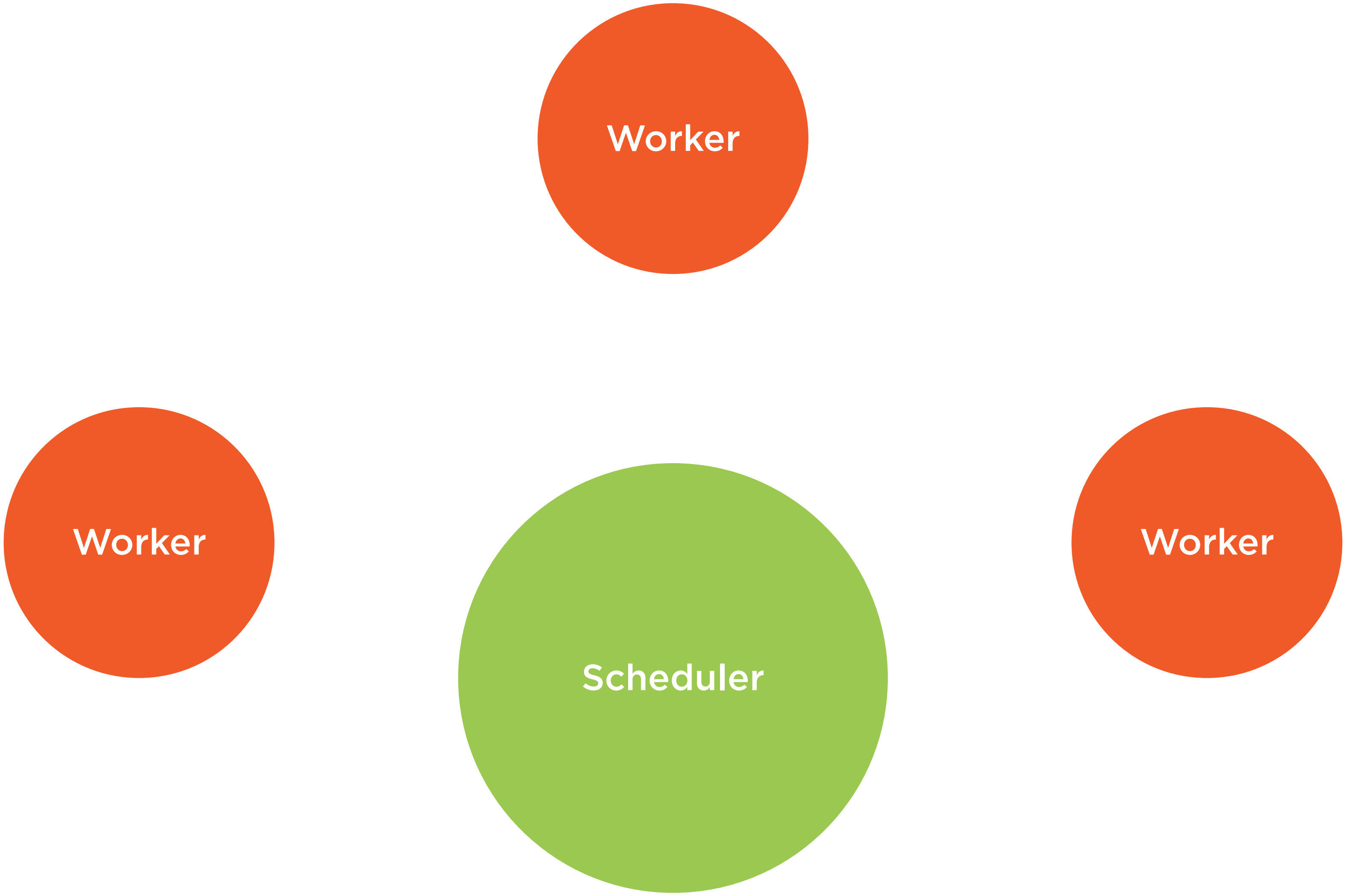
Runtime





Schedulers read and execute task graphs.





Worker

Worker

Scheduler

Worker

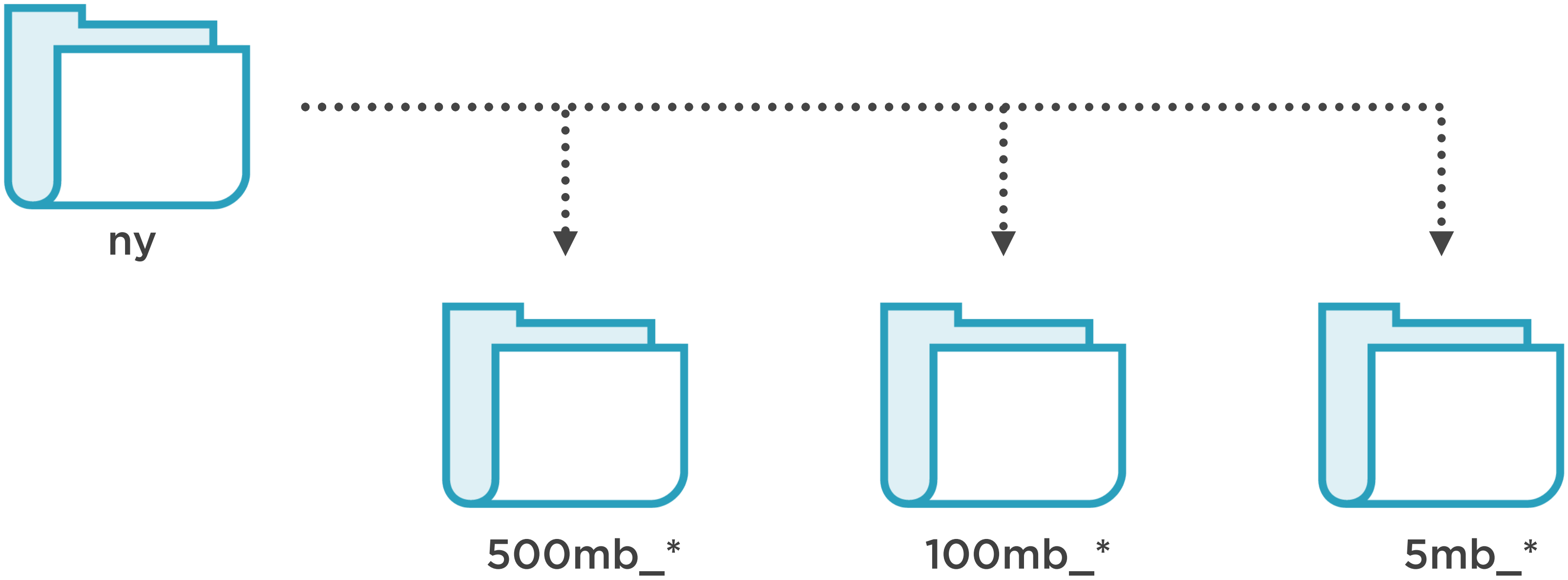
Schedulers

Single-machine

Distributed

Distributed Scheduler Demo

New York Data



Few hundred MBs

Good partition sizing?

Summary

Graph representation

Schedulers

Dashboard