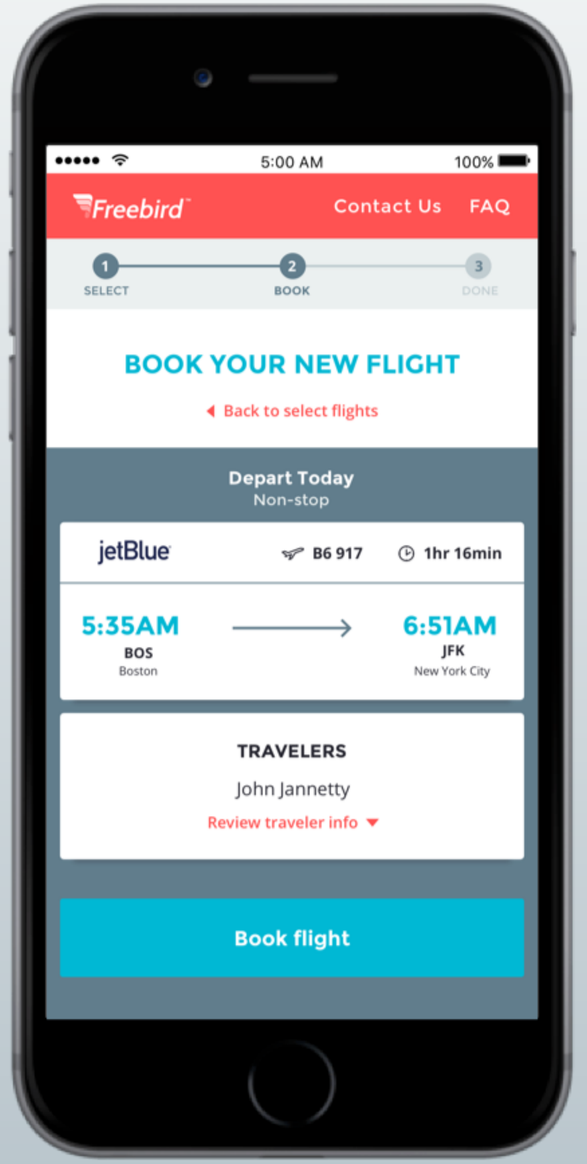
Freebird™

The Goldilocks Problem of Big Data

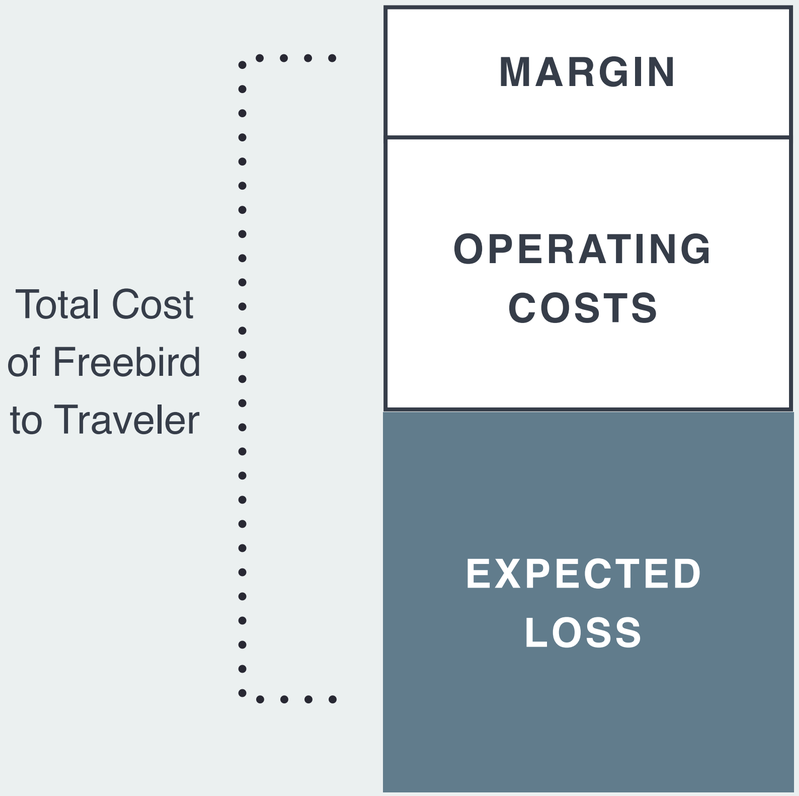
WHAT IS FREEBIRD?

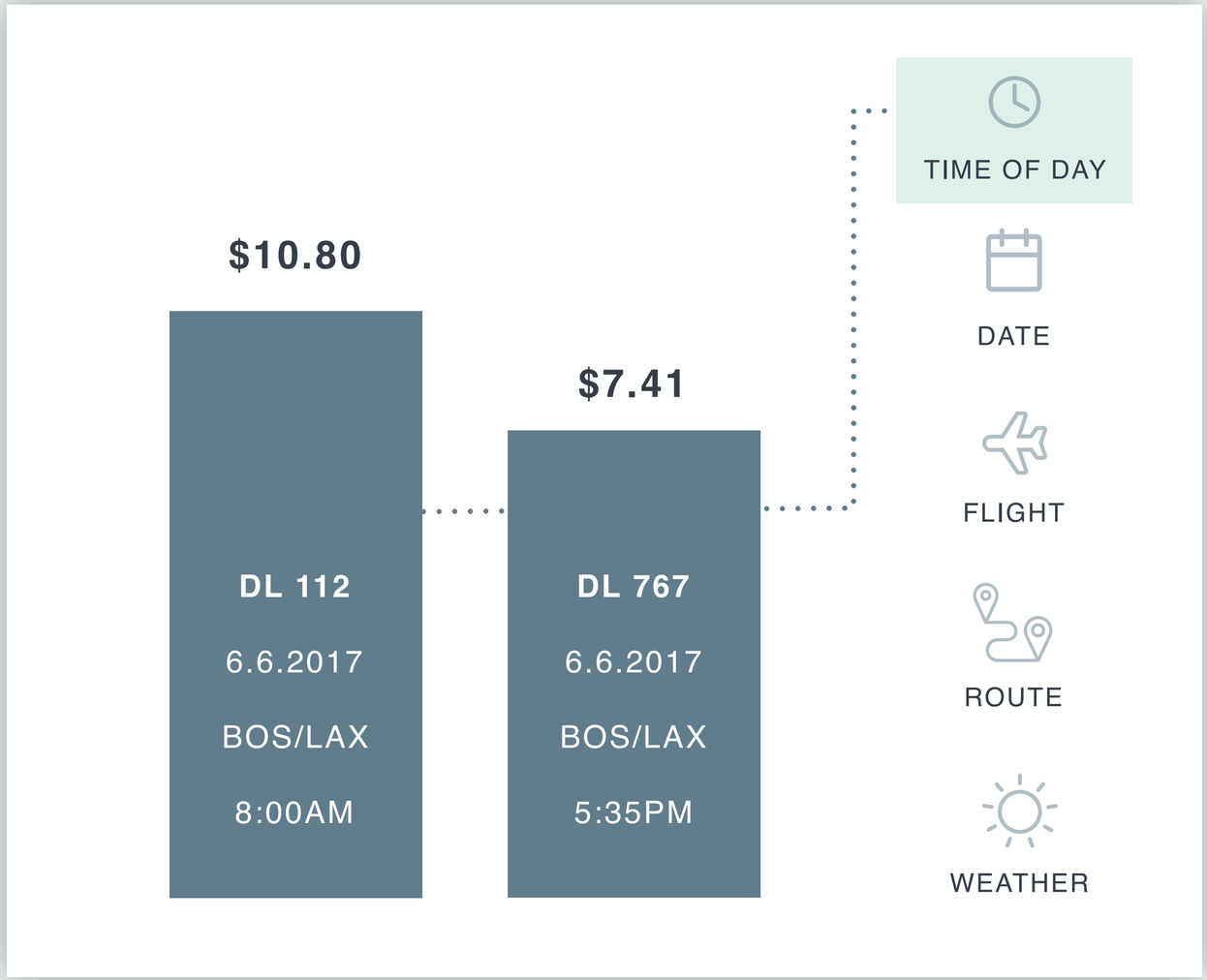
Mobile flight rebooking solution

- Add Freebird before your trip
- If flight is disrupted we send rebooking link
- Rebook in less than 3 taps on any airline for FREE.



WHY BIG DATA?


MARGIN
OPERATING COSTS
EXPECTED LOSS
Total Cost of Freebird to Traveler


$10.80
DL 112
6.6.2017
BOS/LAX
8:00AM
$7.41
DL 767
6.6.2017
BOS/LAX
5:35PM
TIME OF DAY
DATE
FLIGHT
ROUTE
WEATHER

WHAT BIG DATA?

TRIPS

130 TB

FLIGHT STATUS

200 MB/day

WEATHER

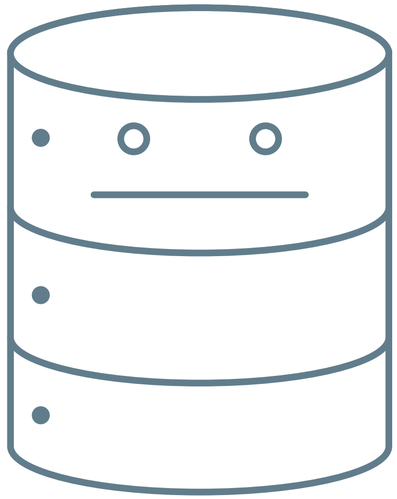
4 GB/day

54KB

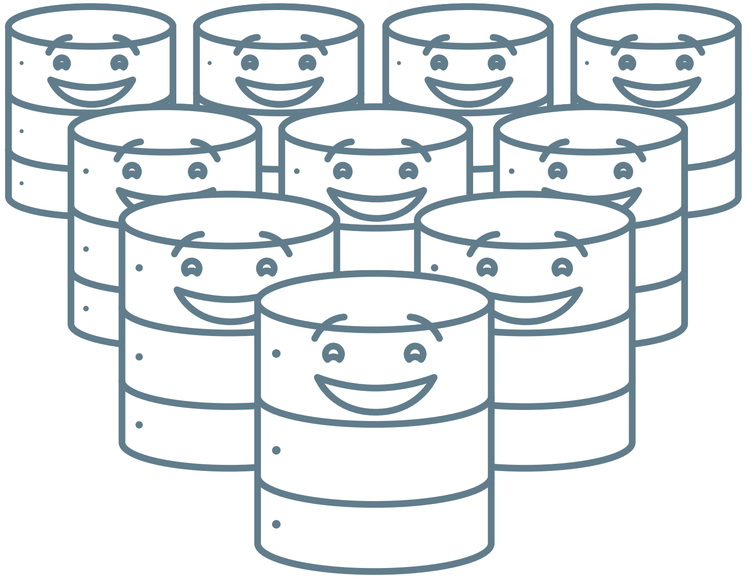
mean file size

SMALL FILES/BIG DATA

6 Months to Query



1 Month to Query



1-6 Hours to Query

THE GOLDILOCKS PROBLEM

FILES TOO SMALL

< 64 MB

Hard drives don’t like jumping from one small file to another

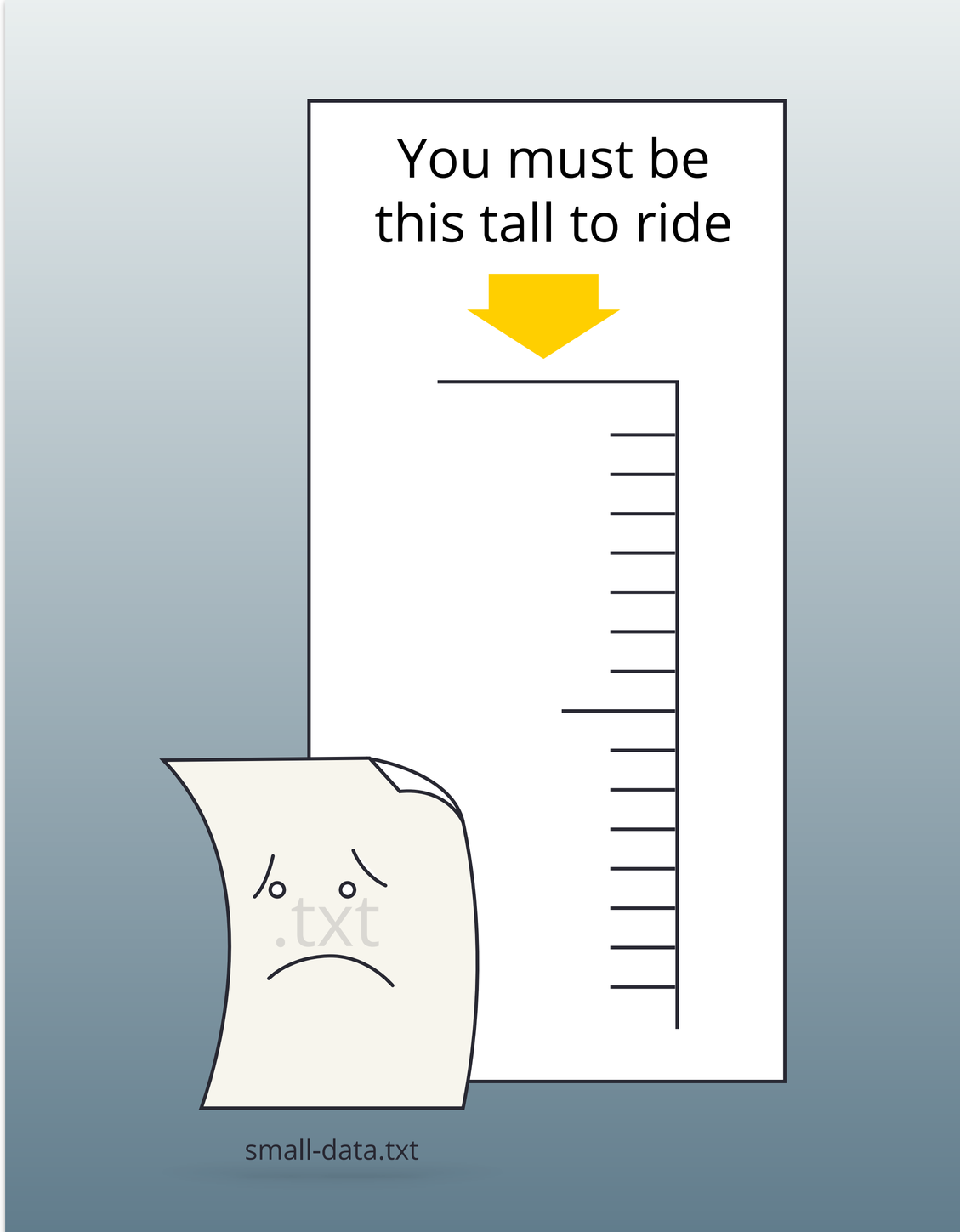
FILES TOO BIG

> 10 GB

Unable to delegate multiple servers to process data in parallel

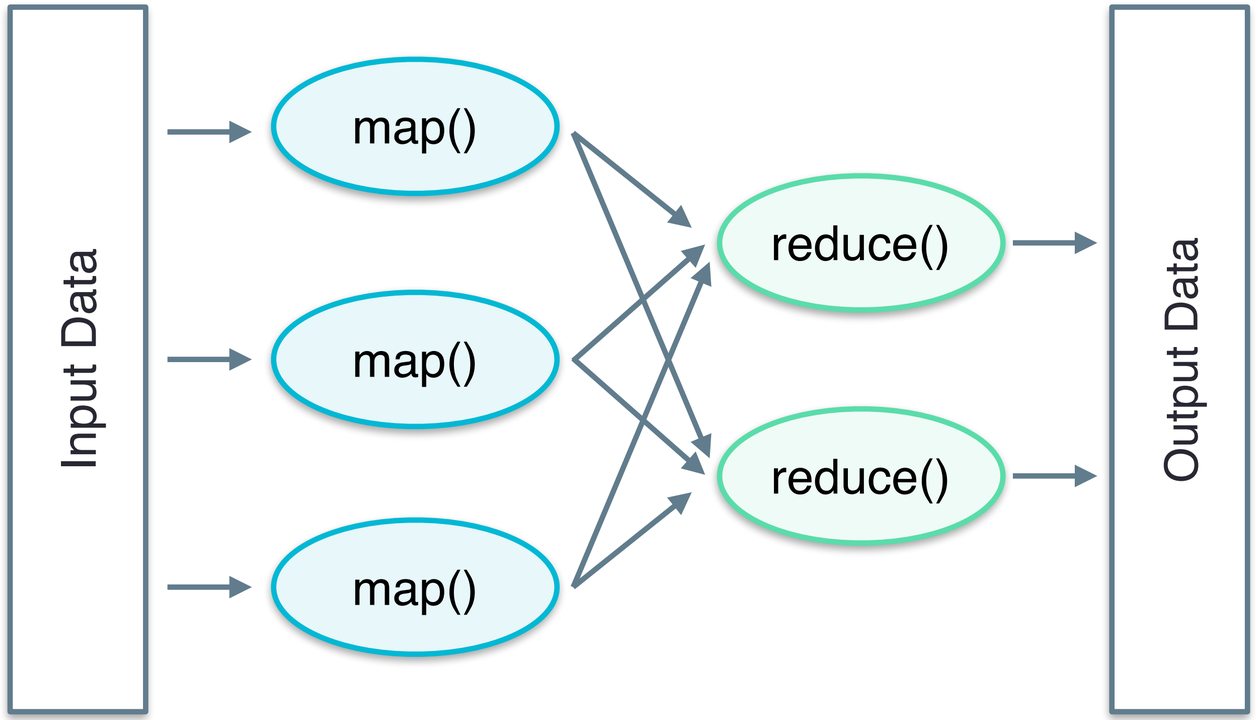
ONE SIDE OF THE PROBLEM

- You want to use big data technologies to process all your BIG data.
- But, big data technologies won’t process your small-file-sized BIG data.



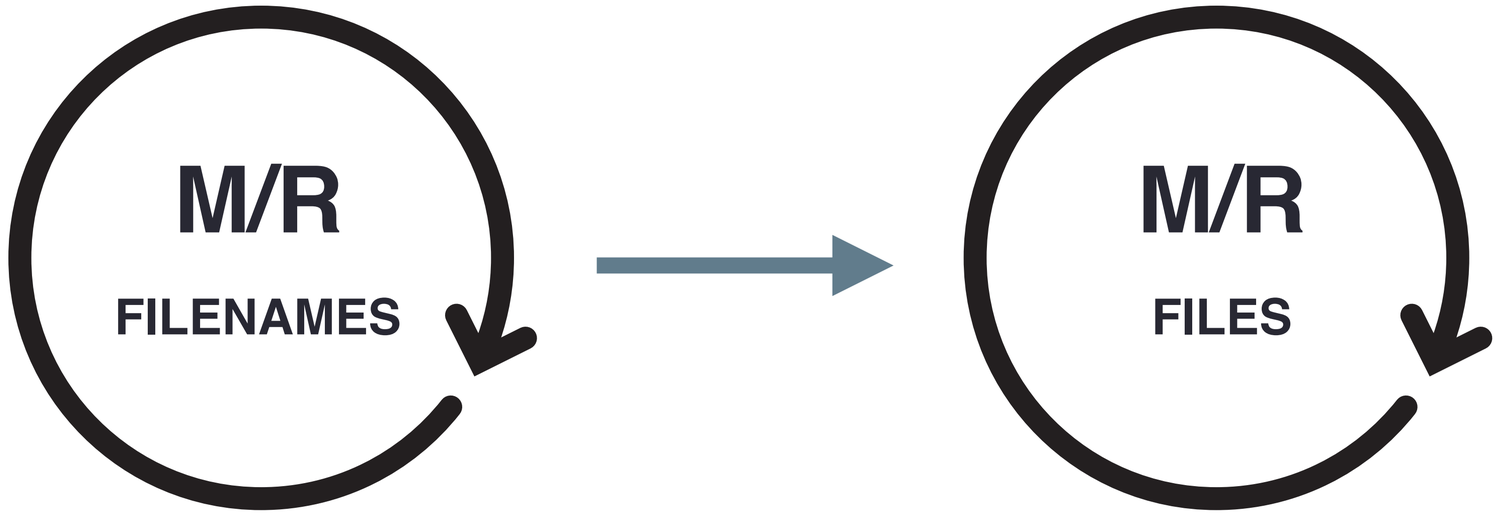
MAP/REDUCE TO THE RESCUE

Map/Reduce is a programming model that allows for tasks to be both **parallelized** and **distributed**.



MAP/REDUCE TO THE RESCUE

Map/Reduce your filenames, then map/reduce files.



TUNING YOUR JOB WILL TAKE TIME

RIGHT FILE SIZE. NO PROBLEM.

Query faster and save money

Happy Data science teem

TensorFlow

TAKEAWAYS

TAKEAWAY 1

Hadoop likes files to be a certain size, not too big or too small.

TAKEAWAY 2

If you **do** have lots of small files, a custom map/reduce job may be needed.

TAKEAWAY 3

Filenames can be big data, too!

Freebird™

Thanks!

FREEBIRD DATA SCIENCE TEAM

(we're hiring!)

TJ Vandal

Max Livingston

Paul Kernfeld

John Russell

Came Piho

Sam Zimmerman
sam@getfreebird.com