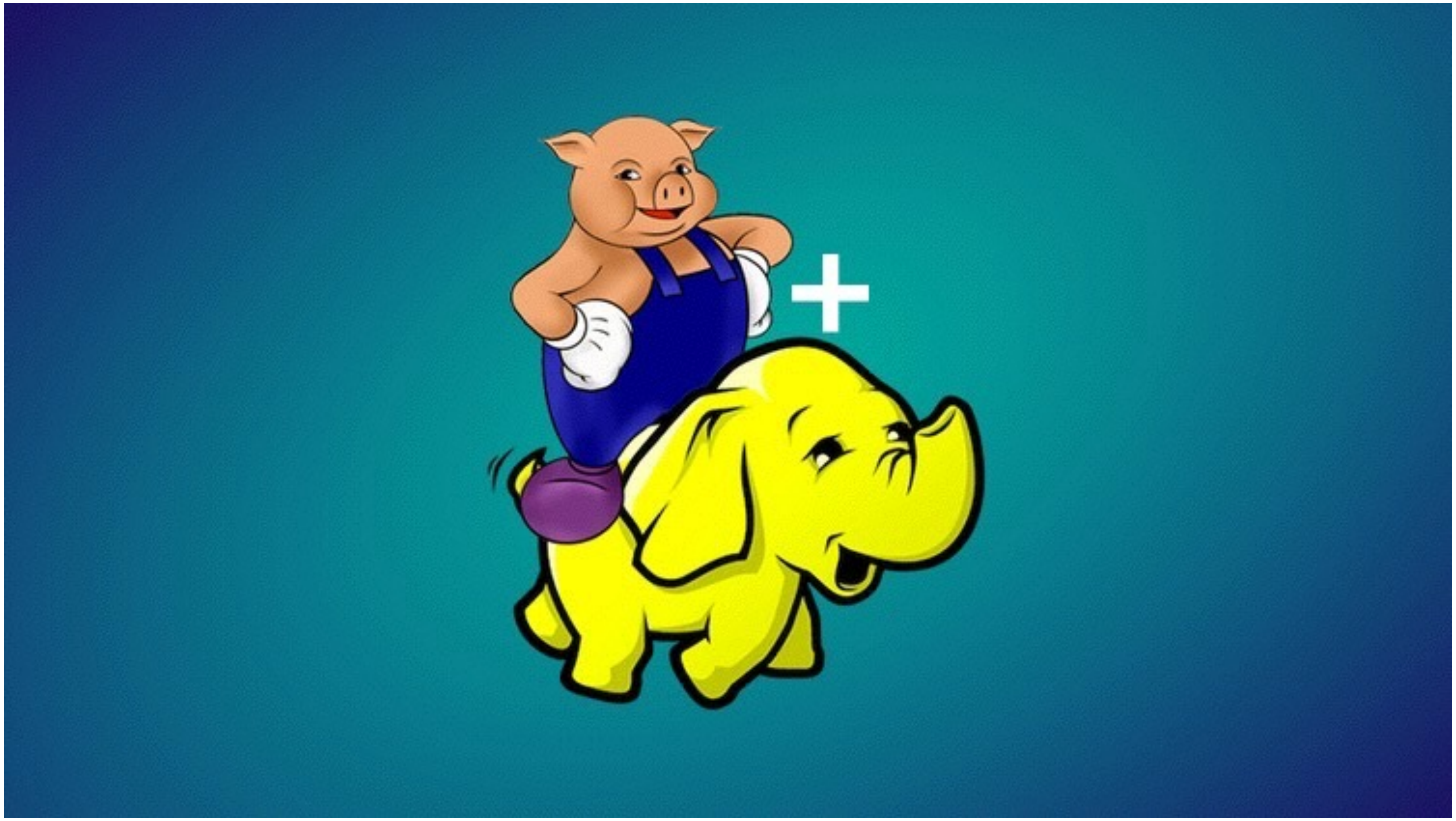
I Never Met a Pig I Didn't Like*

Graeme Ludwig



* The full quote is here. This talk may also contain other pig-related quotes.

What should I cover?

“I have very little idea about Hadoop generally, let alone Pig, and I imagine a few people in the team are in the same boat.

A short presentation on what Hadoop is designed to solve, and a rough idea of how you'd use it without Pig, then a rough idea of how adding Pig helped you would make a great brown bag.”

What am I going to cover?

- Overview of Hadoop & the types of problem it’s designed for
- A Map-Reduce example in Java
- Pig - basic script, UDFs, testing

What is Hadoop?

- A scalable fault-tolerant distributed system for data storage and processing (open source under Apache license)

- Scalable data processing engine

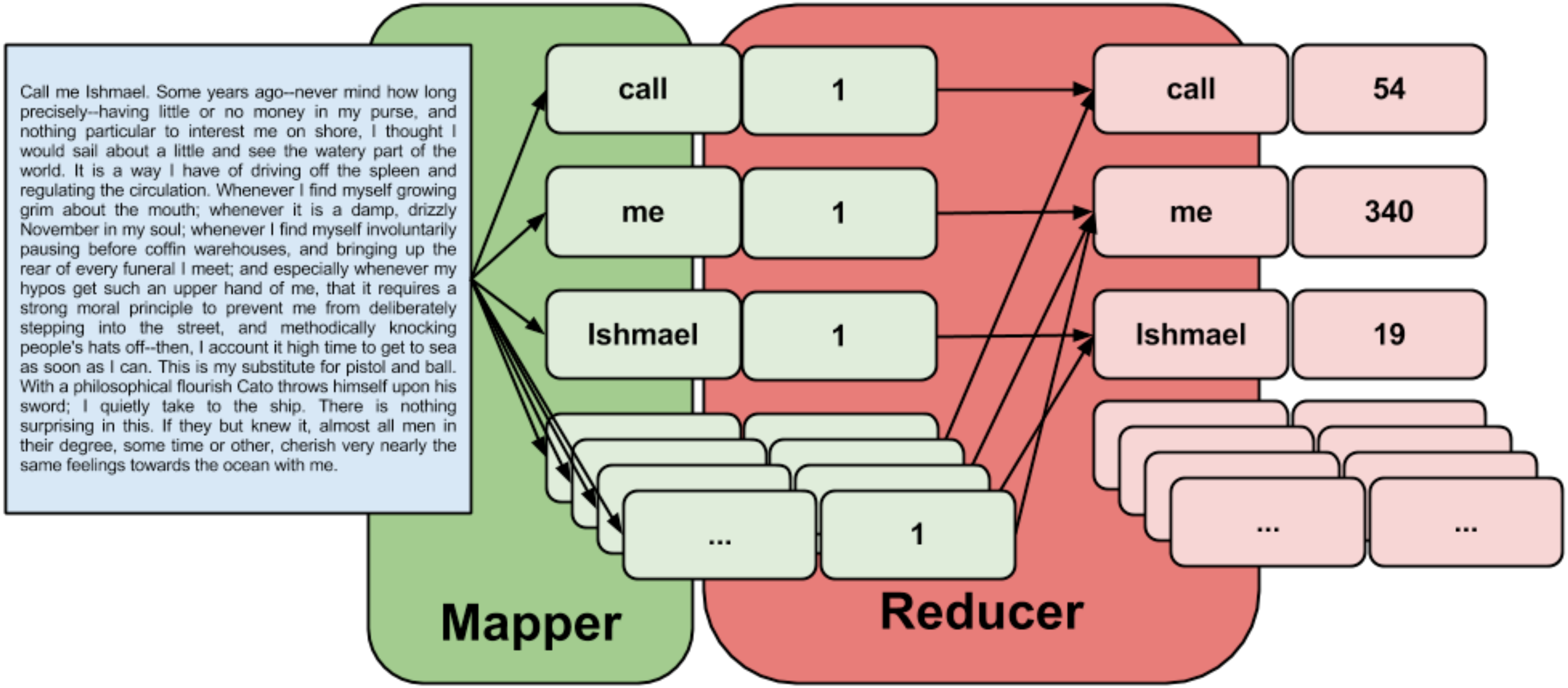
    - Hadoop Distributed File System (self healing, high bandwidth clustered storage)

    - MapReduce: fault tolerant distributed processing

Map Reduce

- A MapReduce program is composed of:
  - a **Map** method that performs filtering and sorting e.g. sorting students by first name into queues, one queue for each name
  - a **Reduce** method that performs a summary operation e.g. counting the number of students in each queue, yielding name frequencies
- **Hadoop** orchestrates the processing by:
  - marshalling the distributed servers
  - running the various tasks in parallel
  - managing all communications and data transfers between the various parts of the system
  - providing for redundancy and fault tolerance

The Hadoop “Hello World”



Call me Ishmael. Some years ago--never mind how long precisely--having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off--then, I account it high time to get to sea as soon as I can. This is my substitute for pistol and ball. With a philosophical flourish Cato throws himself upon his sword; I quietly take to the ship. There is nothing surprising in this. If they but knew it, almost all men in their degree, some time or other, cherish very nearly the same feelings towards the ocean with me.

What Problems Does Hadoop Solve?

- https://www.slideshare.net/cloudera/20100806-cloudera-10-hadoopable-problems-webinar-4931616/7-Summary_10_Common_Hadoopable_Problems

What is Pig?

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

It’s designed for "embarrassingly parallel" data analysis tasks.

OR…

“SQL for Hadoop with power ups”

Example script

```
data = LOAD '../input/example.txt' AS
(line:Chararray);

words = FOREACH data GENERATE
FLATTEN(TOKENIZE(line, ' ')) as word;

grouped = GROUP words by word;

wordcount = FOREACH grouped GENERATE group,
COUNT(words);
```

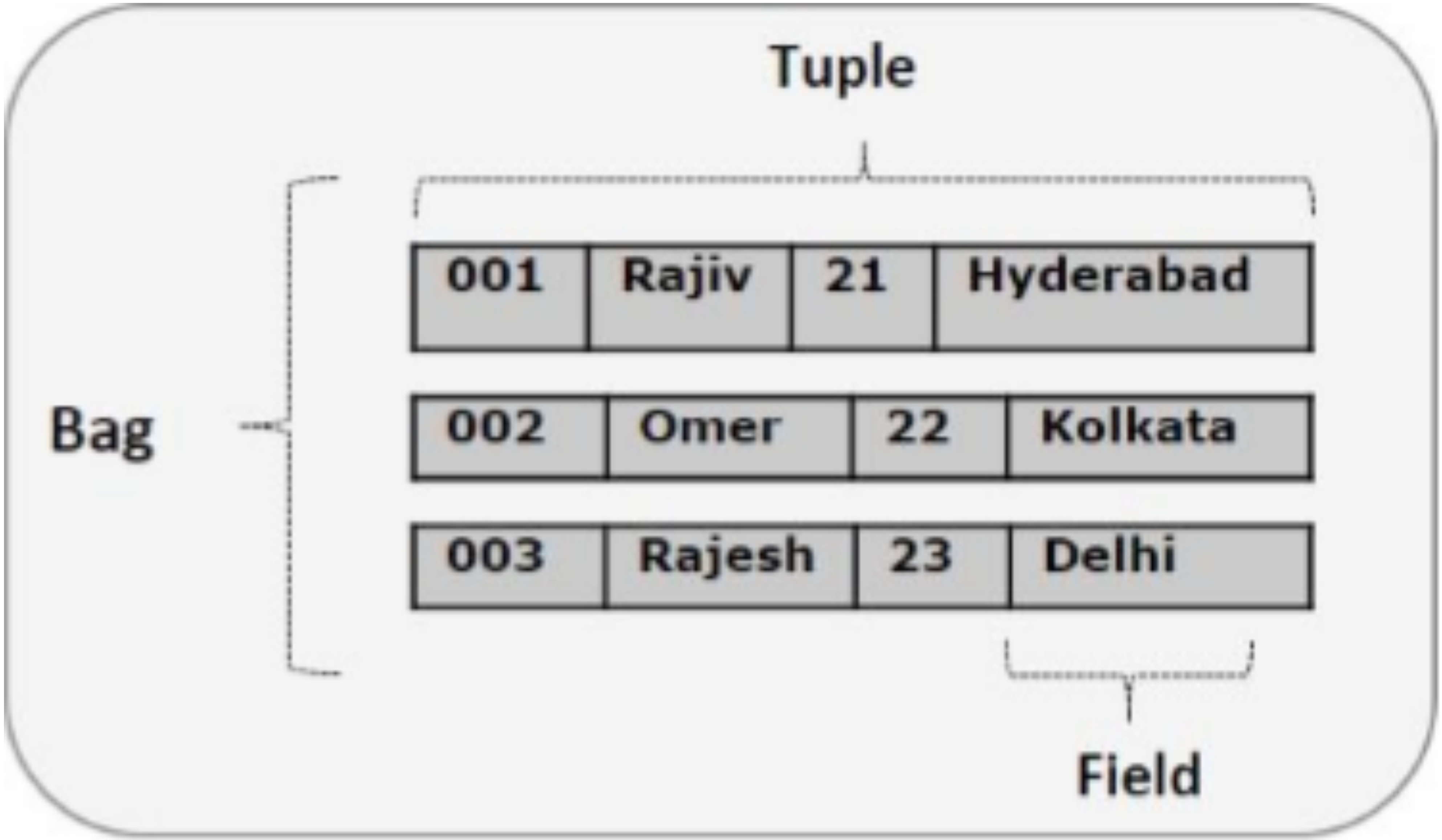
Pig Data Types

- **Atom**
  Single field, stored as string. Can be used as different types e.g. int, long, bytearray…

- **Tuple**
  Ordered set of fields of any type. Similar to a table row.
  e.g. (Raja, 30)

- **Bag**
  Unordered set of tuples. Tuples can have different number and types of fields.
  e.g. {(Raja, 30), (Mohammad, 45)}

Visually…



001	Rajiv	21	Hyderabad
002	Omer	22	Kolkata
003	Rajesh	23	Delhi

UDFs (the power up)

```
data = LOAD '../input/example.txt' AS (line:Chararray);

words = FOREACH data GENERATE
FLATTEN(TOKENIZE(line, ' ')) as word;

grouped = GROUP words by word;

wordcount = FOREACH grouped GENERATE group,
COUNT(words);
```

(See https://pig.apache.org/docs/r0.16.0/udf.html)

How do we test Pig?

- UDFs => Junit, Java, Groovy, Spock
- Pig Scripts => Pig Unit (https://pig.apache.org/docs/r0.8.1/pigunit.html)

The other pig-related quote

I learned long ago, never to wrestle with a pig. You get dirty, and besides, the pig likes it.

George Bernard Shaw