mrjob: part 2

steve johnson - steve@steveasleep.com

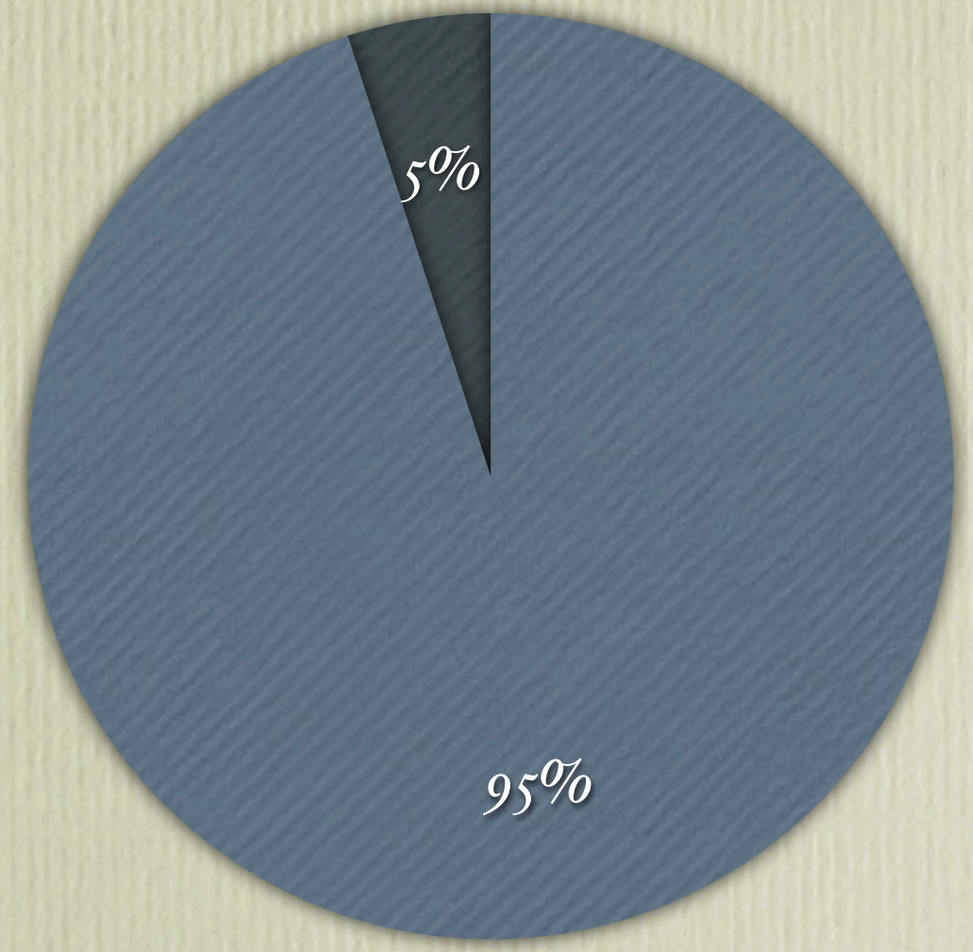
Common friends exercise, explained

What we know now (?)

- What the map step is
- What the reduce step is
- Basic structure of a MRJob class
- Running a job from the command line

Amount of important stuff you know

- Known
- Everything else



What is missing?

- Annoying details
- Specifics of systems (Hadoop, Elastic MapReduce)
- Optimization strategies
- Debugging strategies
- Algorithm intuition

Yelp data installed?

Exercise: review with most unique words (local)

EMR Setup Checklist

1. Get your Amazon codes from your email
2. Get your "access key" and "secret key" from http://aws.amazon.com/account/
3. Create an S3 bucket called "yourname-mr" at https://console.aws.amazon.com/s3/home
4. Put this in ~/.mrjob.conf

```
runners:
  emr:
    aws_access_key_id: <your key ID>
    aws_secret_access_key: <your secret>
    s3_log_uri: s3://yourname-mr/logs/
    s3_scratch_uri: s3://yourname-mr/tmp/
```

Exercise: review with most unique words (EMR)

Exercise: MapReduce grep (command line arguments)

Exercise: MapReduce grep (command line arguments)

Exercise: MapReduce grep using mapper_cmd()

Homework

- There is an input file at s3://sjohnson-public/yelp_reviews.json
- For each user, output the user ID and the highest number of consecutive days they checked in (so if I check in Mon Tues, Fri, Sat, Sun, then it would be 3).
- Test with data/yelp/reviews_100.json
- Must run with S3 input on EMR also
- Bonus: also print the user name