Map Reduce and Streaming Calculations

Giri Iyengar

Cornell University

gi43@cornell.edu

April 16, 2018

Agenda for the week

- Map-Reduce
- Poisson resampling
- Streaming Calculations
  1. Reservoir Sampling
  2. Storing Items in Sets
  3. Counting in single pass
  4. Frequent Items in a stream
  5. Estimating CDF/PDF in streaming mode
- Background Reading

Overview

Hello World in Map Reduce

Counting words in some text

But I must explain to you how all this mistaken idea of denouncing pleasure and praising pain was born and I will give you a complete account of the system.

- Step 0: Parse text word by word.
- Step 1: Emit one word at a time
- Step 2: Group same words together
- Step 3: Count occurrence of each word

Word Count - Emit

```
But	1
I	1
must	1
explain	1
to	1
you	1
how	1
all	1
this	1
mistaken	1
idea	1
of	1
denouncing	1
pleasure	1
```

Word Count - Emit

```
and             1
praising             1
pain             1
was             1
born             1
and             1
I             1
will             1
give             1
you             1
a             1
complete             1
account             1
of             1
the             1
system.             1
```

Word Count - Sort and Group

```
But                  1
I                  1
I                  1
a                  1
account                  1
all                  1
and                  1
and                  1
born                  1
complete                  1
denouncing                  1
explain                  1
give                  1
how                  1
idea                  1
mistaken                  1
```

Word Count - Sort and Group

```
must        1
of      1
of      1
pain        1
pleasure        1
praising        1
system.     1
the     1
this        1
to      1
was     1
will        1
you     1
you     1
```

Word Count - Aggregate

```
But             1
I             2
a             1
account             1
all             1
and             2
born             1
complete             1
denouncing             1
explain             1
give             1
how             1
idea             1
mistaken             1
```

Map Reduce

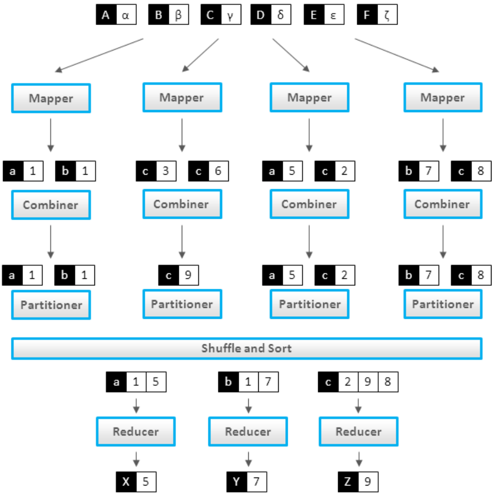


Figure: Source: highlyscalable.com blog

MR Demo

Map Reduce Examples

- Word Count

Map Reduce Examples

- Word Count
- Unique Count

Map Reduce Examples

- Word Count
- Unique Count
- Total Sales, Average Sales by Customer

Map Reduce Examples

- Word Count
- Unique Count
- Total Sales, Average Sales by Customer
- Click-Through-Rate of Advertising Campaigns

Map Reduce Examples

- Word Count
- Unique Count
- Total Sales, Average Sales by Customer
- Click-Through-Rate of Advertising Campaigns
- Little Bag of Bootstraps to Build Models

HDFS: Hadoop Distributed File System

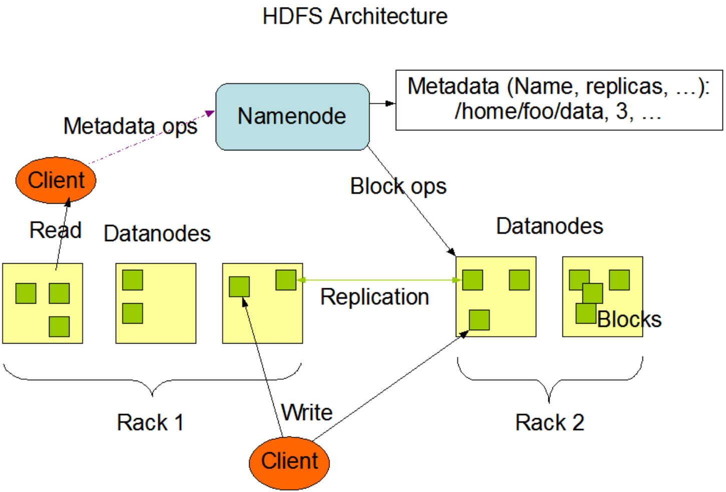


Figure: Source: Apache Foundation

Hadoop Ecosystem

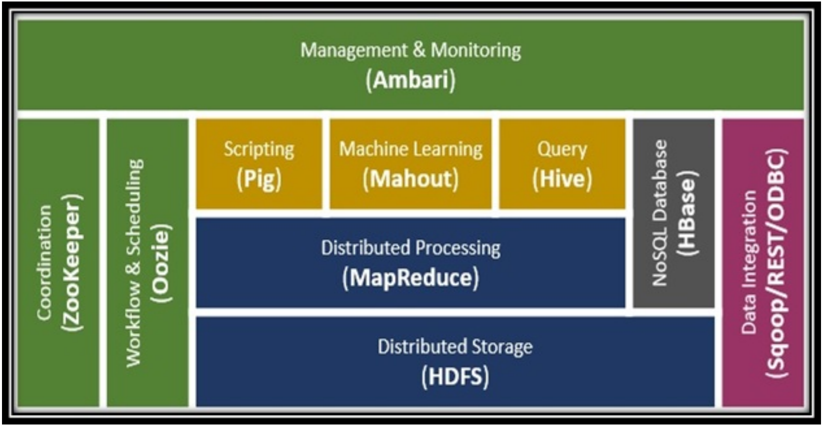


Figure: Source: Apache Foundation

Overview

Poisson Resampling for Efficient Bootstrapping

Bootstrapping Big Data

- Bootstrap is not efficient when dealing with Big Data $\approx 63.2\%$ of data gets resampled per bootstrap
- Need several bootstrap samples depending on what you are trying to estimate (e.g. std err or percentiles)
- Little Bag of Bootstraps is one technique (saw last week)
- Poisson Resampling is another technique

Poisson Resampling

Motivation

Let's start with a tiny sample $\{1.5, 2.5, 3.5, 4.5\}$ and do bootstrap

Sample 1	$\{1.5, 1.5, 3.5, 2.5\}$
Sample 2	$\{2.5, 1.5, 3.5, 2.5\}$
Sample 3	$\{3.5, 4.5, 4.5, 4.5\}$

We can actually describe this in terms of sample counts

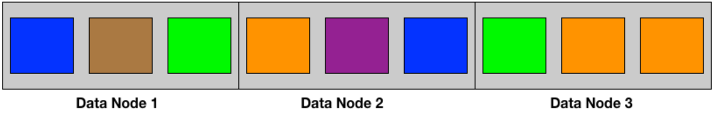
Sample 1	$\{2, 1, 1, 0\}$
Sample 2	$\{1, 2, 1, 0\}$
Sample 3	$\{0, 0, 1, 3\}$

These counts follow a $Multinomial(4, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ distribution. Generally, $Multinomial(n, \frac{1}{n}, \ldots, \frac{1}{n})$

Poisson Resampling

Big Data Problem

Not all data resides in the same place. Data is distributed. Also, in streaming cases, we may not even know n in advance



Can we *approximate* bootstrapping without bringing all the data to one place?

Poisson Resampling

Approximation

What if we independently sample each data point using a $Binomial(n, \frac{1}{n})$? All sampling can be done in parallel. For large n, this is *close-enough* to multinomial sampling that it doesn't matter in practice. But, we still need to know n in advance!

Poisson Distribution

$$\lim_{n \to \infty} Binomial(n, \frac{1}{n}) = Poisson(1)$$

$Poisson(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ doesn't need to know n!

Poisson Resampling in Action

Poisson Resampling in Map-Reduce

- Independently sample in your map task using $Poisson(1)$
- Emit those *k* samples
- Reducers get independent datasets, run their aggregations, and return back results
  - Aggregations could be statistics, or even entire models!


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
arigatō
dakujem
grazie
terima kasih
감사합니다
ありがとう
мерси
tapadh leat
dank je
takk
taiku
sagolun
hvala
mauruuru
Баярлалаа
рахмат
хвала
εὐχαριστώ