

# Contents

<b>Big Data Computing</b>	<b>1</b>
First Lecture . . . . .	1
Second Lecture . . . . .	1
What's Big Data? . . . . .	1
Scaling . . . . .	1
Network Bottlenecks . . . . .	2



# Big Data Computing

## First Lecture

Classroom: phdlyrl

## Second Lecture

### What's Big Data?

It is sometimes used as a buzzword, but it describes an actual phenomenon.

5 V's:

1. Value: Extracting knowledge from data.
2. Volume: The amount of data. (Measuring in terabytes, petabytes, exabytes)
3. Variety: Different formats, structured (SQL tables) and Unstructured (text, images, videos).
4. Velocity: The speed at which data is generated. (usually in real-time, you can't parse it all)
5. Veracity: The quality of the data. (Is it reliable?)

### Scaling

When scaling, you have two options:

- Scaling *up*, you buy better hardware for your monolithic machine, however, moore's law is slowing down, and you will get diminishing returns, in general, you will not be able to keep up with increases in Volume/Velocity.
- Scaling *out*, you buy more machines, and distribute the workload, this is the way to go for big data, since this allows for exponential growth.

Scaling *out*, or *horizontal scaling*, also allows you for more flexibility, as once you deal with the orchestration of the workload, you can dynamically increase/reduce the number of allocated machines depending on a rolling cost/benefit analysis.

## Network Bottlenecks

A bottleneck during a request is generated when:

- There is a skew in the ratio of communication/computation
- There is a non-overlappable part of communication/computation