PHÂN TÍCH DỮ LIỆU LỚN

GV: ThS. Nguyễn Thế Bảo
Chương trình HTTT&AI, Viện KTCN
Email: baont@tdmu.ed.vn
Zalo: 0937.068.474

NỘI DUNG LÝ THUYẾT

CHƯƠNG 1: TỔNG QUAN

CHƯƠNG 2: XỬ LÝ DỮ LIỆU LỚN TRÊN MAPREDUCE

CHƯƠNG 3: TÌM KIẾM TƯƠNG TỰ

CHƯƠNG 4: KHAI THÁC TẬP PHỔ BIẾN

CHƯƠNG 5: GOM CỤM

CHƯƠNG 6: MÔ HÌNH KHUYẾN NGHỊ

CHƯƠNG 7: PHÂN TÍCH MẠNG XÃ HỘI

NỘI DUNG THỰC HÀNH & BÀI TẬP

TH1: Cài đặt Hadoop, Spark.

TH2: Thực thi các hàm thông dụng của Spark

TH3: Cài đặt example trên Spark (word count,…)

TH4: Bài tập KNN, Kmeans

TH5: Bài tập A-Priori

TH6: Bài tập page rank

PHÂN BỐ TUẦN

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LT C1	LT C2	LT C2	LT C2	LT C3	BT C3	LT C4	LT C5	BT C5	LT C6	BT C6	LT C7	BT C7	ÔN TẬP	BÁO CÁO BÀI GK

ĐÁNH GIÁ

- GIỮA KỲ: 50%
  - ĐIỂM DANH: 10%
  - BÀI THỰC HÀNH: 15%
  - BÁO CÁO CÁ NHÂN: 25%
- CUỐI KỲ: 50%
  - TIỂU LUẬN: SINH VIÊN THỰC HIỆN THEO NHÓM,

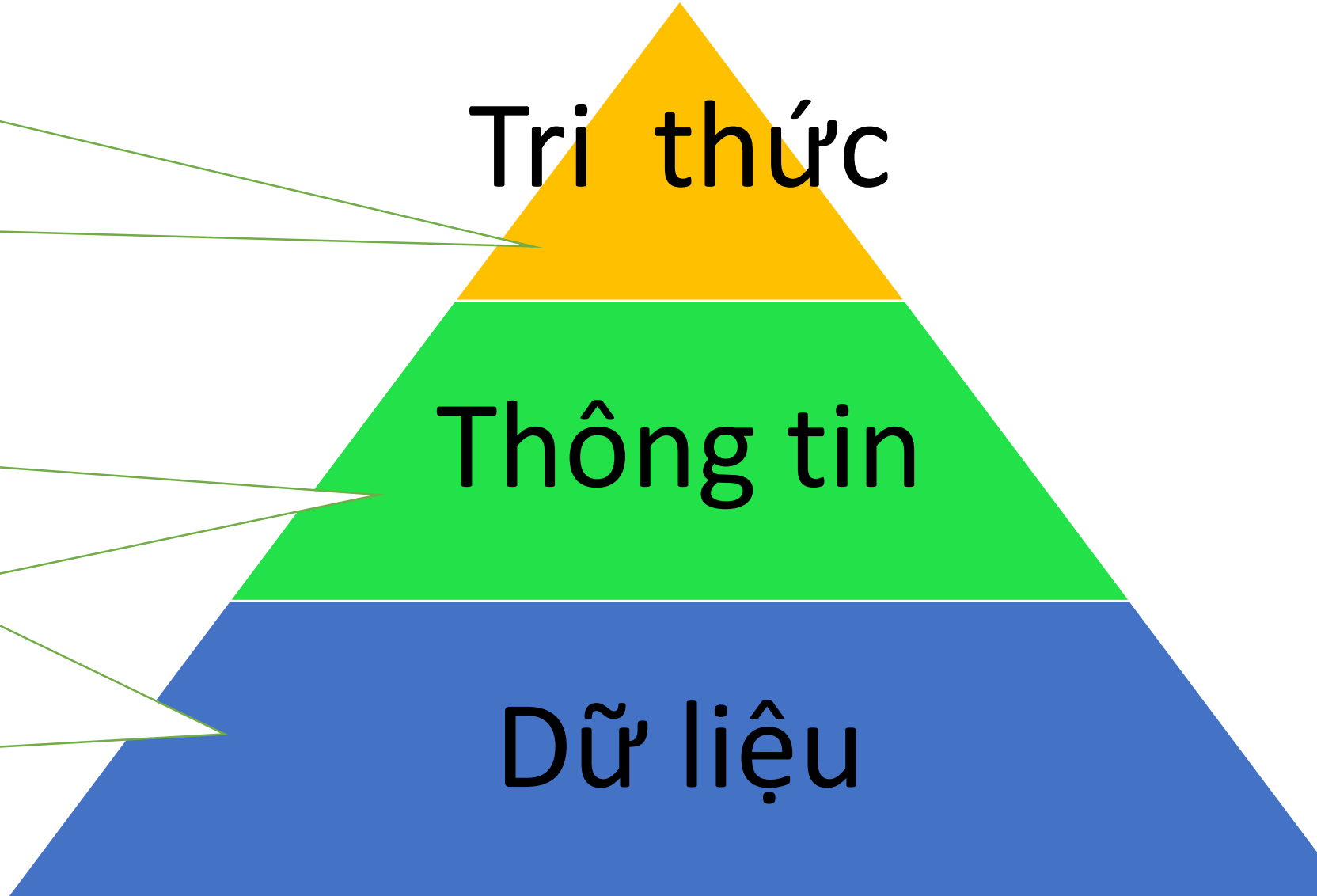
CHƯƠNG 1: TỔNG QUAN

1.1. Giới thiệu

Là thông tin tích hợp như quan hệ giữa các sự kiện, giữa các thông tin…thu được qua quá trình nhận thức, phát hiện hoặc học tập

côt lõi đặc trưng cho dữ liệu

diễn bằng các bits, các con số hoặc ký hiệu.



CHƯƠNG 1: TỔNG QUAN

1.1. Giới thiệu

- Đơn vị đo của dữ liệu:

B (byte),

KB (Kilobyte),

MB (Megabyte),

GB (gigabyte),

TB (terabyte),

PB (petabyte)

EB (Exabyte)

ZB (Zetabyte)

….

CHƯƠNG 1: TỔNG QUAN

1.2. Dữ liệu lớn (BigData) là gì?

- Quan điểm 1: dữ liệu lớn là tập dữ liệu rất lớn hoặc rất phức tạp vượt quá khả năng xử lý của các kỹ thuật truyền thống.

CHƯƠNG 1: TỔNG QUAN

1.2. Dữ liệu lớn (BigData) là gì?

- Quan điểm 2: 4V

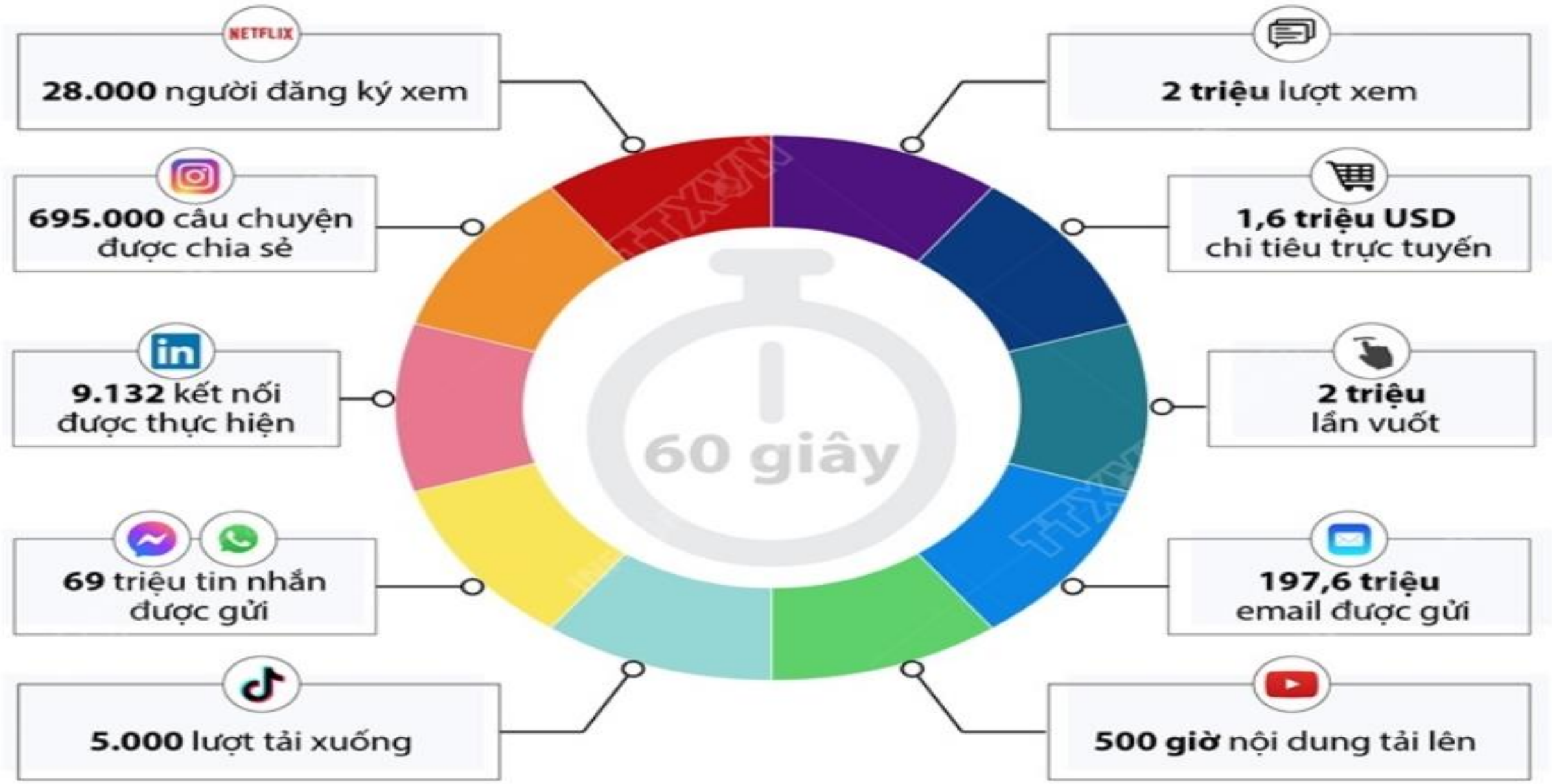
+ Volume: Dung lượng lớn (về số lượng và số thuộc tính)

+ Velocity: dữ liệu không có đích.

+ Variety: tính đa dạng (dữ liệu có nhiều cấu trúc khác nhau)

+ Veracity: độ tin cậy của dữ liệu

LƯỢNG DỮ LIỆU ƯỚC TÍNH ĐƯỢC TẠO TRÊN INTERNET TRONG MỘT PHÚT



Nguồn: Lori Lewis; AllAccess; Statista

1 internet minute

- $ 400 M sales on Alibaba
- 439, 000 page views on Wikipedia
- 194, 000 apps downloaded
- 31, 700 hours of music played on Pandora
- 38, 000 photographs uploaded to Instagram
- 4.1 Million searches on Google
- 139, 000 hours of video watched on Youtube
- 10 million ads displayed
- 3.3 million shares on Facebook

Each of these activities generates

DATA

Alibaba
Wikipedia
Pandora
Instagram
Google
Youtube
Facebook

These companies and others are collecting

PetaBytes of data every minute

CHƯƠNG 1: TỔNG QUAN

1.5. Thu thập dữ liệu lớn

- Nguyên nhân có thể thu thập được dữ liệu lớn ?

Giá thiết bị lưu trữ đã giảm rất nhiều trong nhiều năm trở lại đây.

- Thu thập được dữ liệu lớn để làm gì ?

Vì có thể kiếm tiền từ đó

Everything is personalized

Product Recommendations on Amazon,
Newsfeed on Facebook,
Homepage on Netflix

Ads, Offers, Promotions just for you!

Really cool products can be built

Google Maps,
Apple Siri

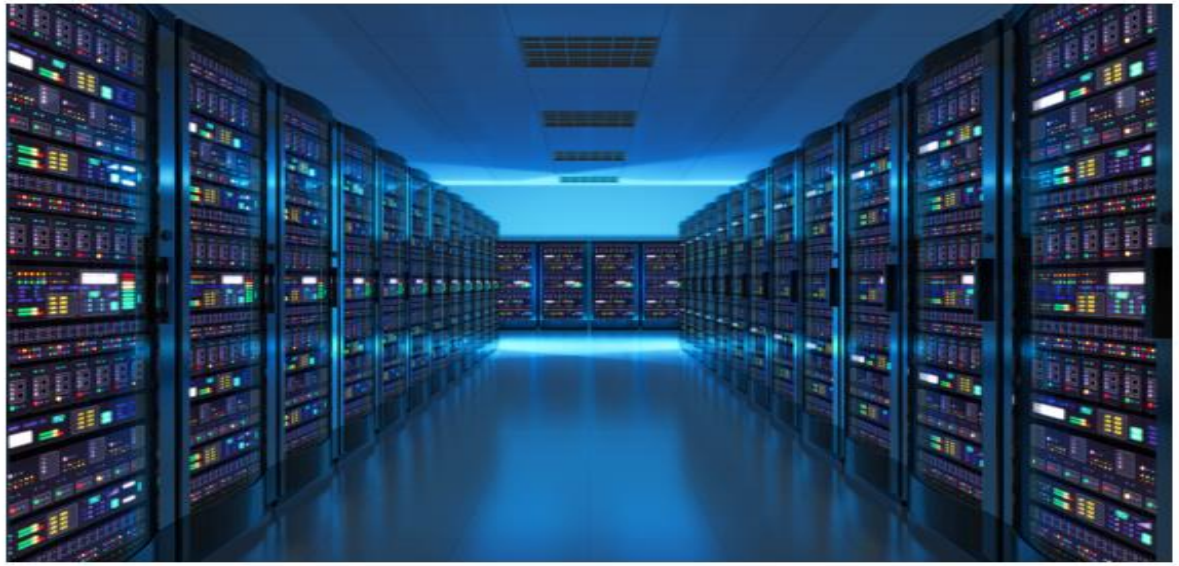
CHƯƠNG 1: TỔNG QUAN

1.6. Lưu trữ dữ liệu lớn

Huge Data Centers

covering 100s of acres

with millions of servers



CHƯƠNG 1: TỔNG QUAN

1.6. Lưu trữ dữ liệu lớn

Huge Data Centers

with millions

of servers

running

sophisticated

proprietary

software

to process

TBs/PBs of

data

Thực thi các phần mềm độc quyền phức tạp **(Hadoop, Spark, Hive,…)** để xử lý dữ liệu lớn

CHƯƠNG 1: TỔNG QUAN

1.7. Truyền dữ liệu lớn

You work for an e-commerce startup

You collect 1 TB worth of weblogs everyday

At the end of the day you want to publish a report on traffic for the day

Phương án?

CHƯƠNG 1: TỔNG QUAN

1.7. Truyền dữ liệu lớn

Khả thi?

Option # 1

Use a single powerful server

1 TB hard disk drive (minimum)

1. Vấn đề về dung lượng ổ cứng lưu trữ
2. Thời gian truyền dẫn

CHƯƠNG 1: TỔNG QUAN

1.7. Truyền dữ liệu lớn

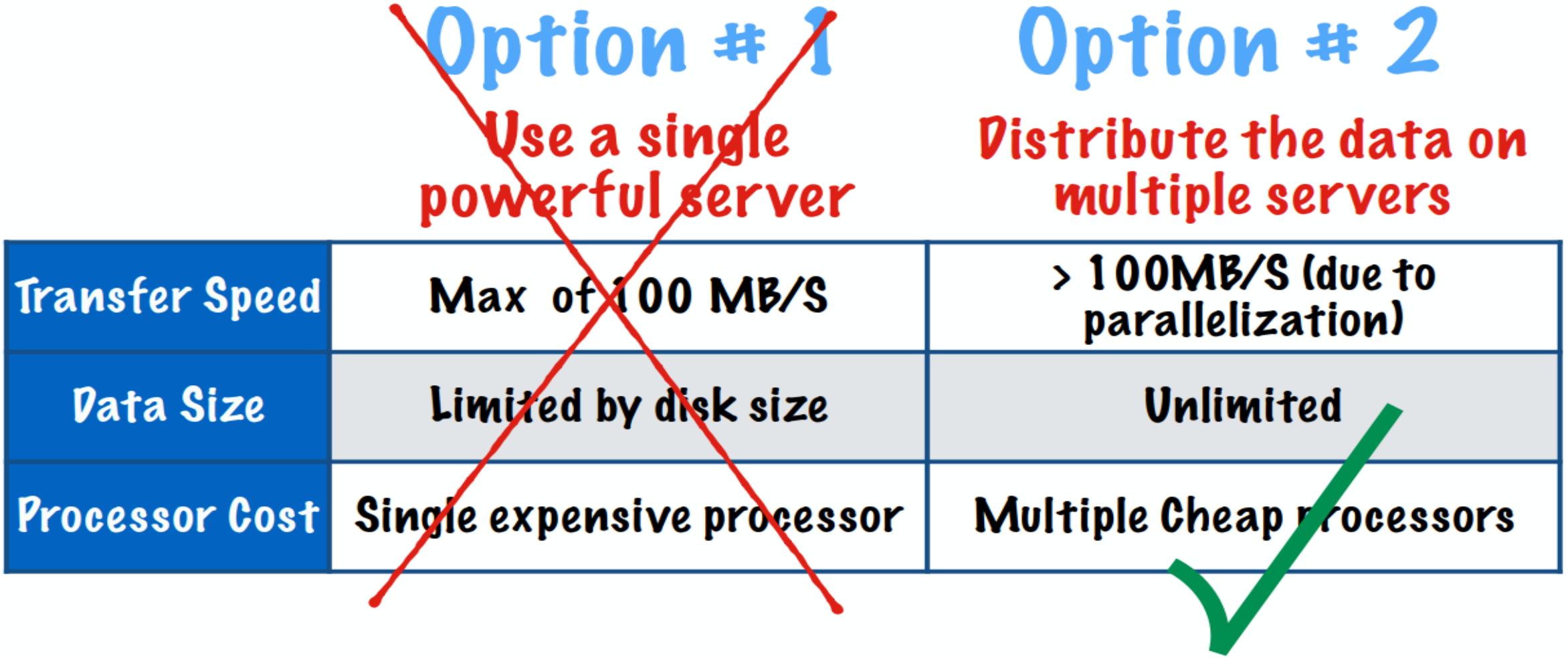
Option # 2

Distribute the data on multiple servers

1. Vấn đề về dung lượng ổ cứng lưu trữ
2. Thời gian truyền dẫn

CHƯƠNG 1: TỔNG QUAN

1.7. Truyền dữ liệu lớn



How do we distribute and process data on multiple servers?