The background of the slide is a dense, abstract composition of three-dimensional numbers. The numbers, including digits 0 through 9, are rendered in a light blue color with a subtle gradient and are positioned at various angles and heights, creating a sense of depth and movement. They appear to be floating or stacked, with some numbers partially obscured by others, giving the overall image a complex, data-driven aesthetic.

Vaex: a faster pandas alternative

Hao Cheng, Jerry Lin,
Justin Mason, Caitlyn Yen,
Elsa Yen



A faster pandas alternative

Typical data analysis workflow involves pandas

Can pandas keep up with this growing trend?

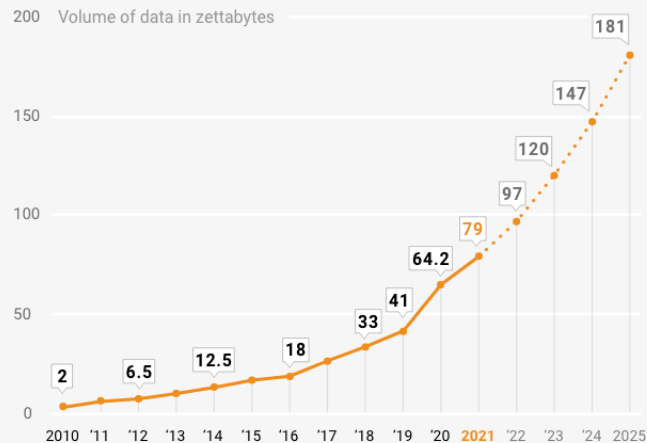
Faster alternatives?



Volume of data created, captured, copied, and consumed worldwide



The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



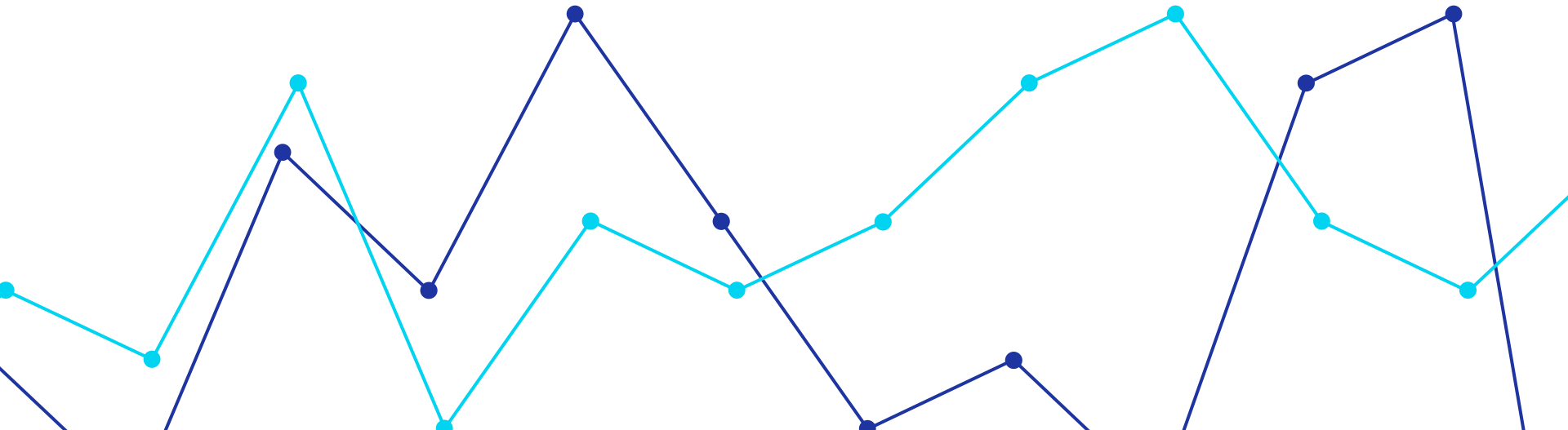
Source: statista.com





- ◇ Lazy
 - ◇ Perform operations only when called
- ◇ Out-of-core
 - ◇ Process data too large to fit in computer's main memory
- ◇ Scale
 - ◇ Datasets as big as your hard drive
- ◇ Performance
 - ◇ 10 billion rows/second
- ◇ Plot
 - ◇ Native support, one-line implementation
- ◇ Virtual columns
 - ◇ Does not take up any memory, computed when needed

Can vaex outperform pandas? SQL?



Strategy



Tools

Pandas

Vaex

SQL



File Sizes

1 GB

5 GB

10 GB



Performance

Data loading

Data cleaning

Plotting

Results



Ease of
Adoption

Scaling
Ability

Scaling
Strategy

 **vaex**



100GB+

Lazy Loading

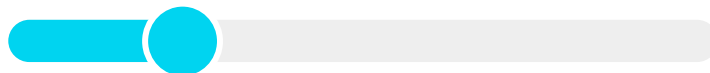
 **dask**



1TB+

Clusters

RAPIDS



100GB+

GPUs

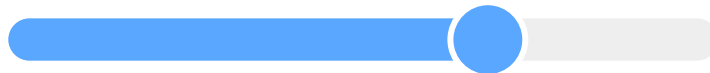
 **MODIN**



10GB+

Clusters

 **RAY**



1TB+

Clusters

Is vaex a pandas killer?



- ◆ Size range: 5-100+ GB
- ◆ Routine Tasks: data cleaning, plotting, basic analysis
- ◆ Pandas to vaex: easy learning curve
- ◆ Advance tasks: look elsewhere



- ◆ Size range: < 1 GB
- ◆ Still the go-to option for most tasks

Questions?

Vaex plot of taxi pickup
locations in New York City

107 GB file
1,173,057,927 rows
0.0233 seconds

