



SOFTWARE DEVELOPMENT CONFERENCE

# YOW! LONDON 2022



# GOTO Guide App

- Download the app
- Ask questions
- Rate sessions



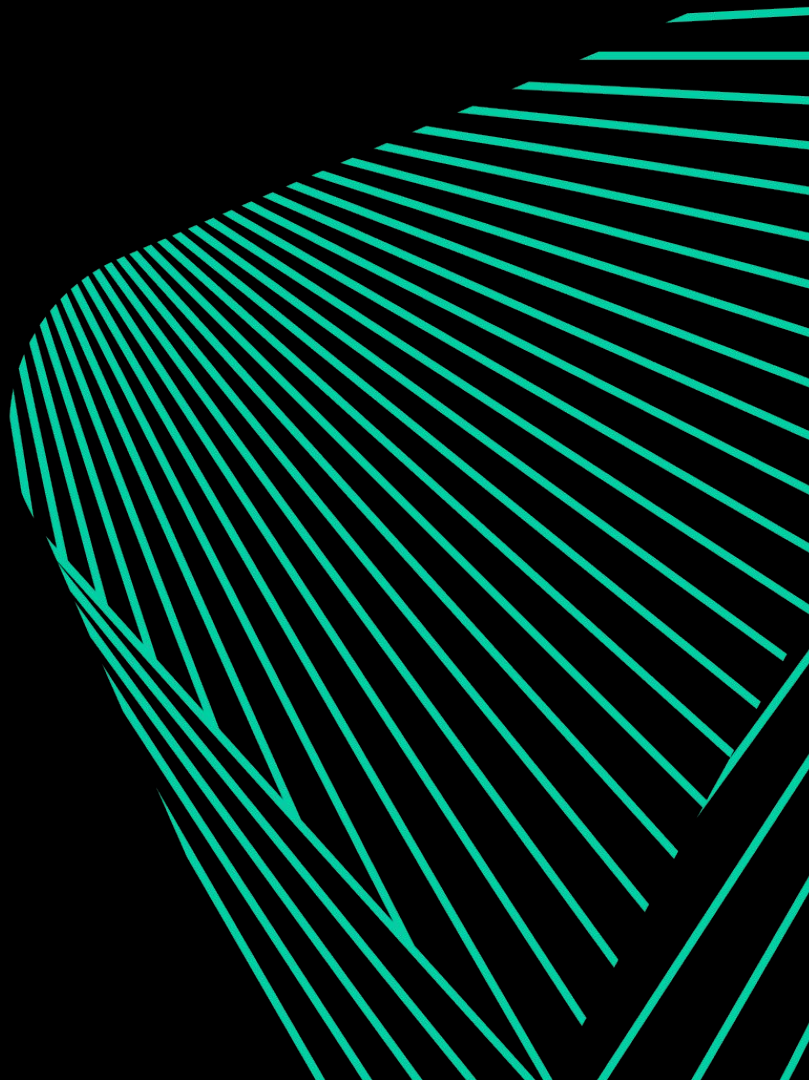
**“Empty your cup so that it may be filled; become devoid to gain totality.”**

**— Bruce Lee**



# Most Data Stacks Aren't Fit For Purpose: The World's Worst Kept Secret

Adam Jennings - Senior Solutions Architect  
Maria Hedenborg - Marketing Director



DATA



SORTED



ARRANGED



PRESENTED  
VISUALLY



EXPLAINED  
WITH A STORY

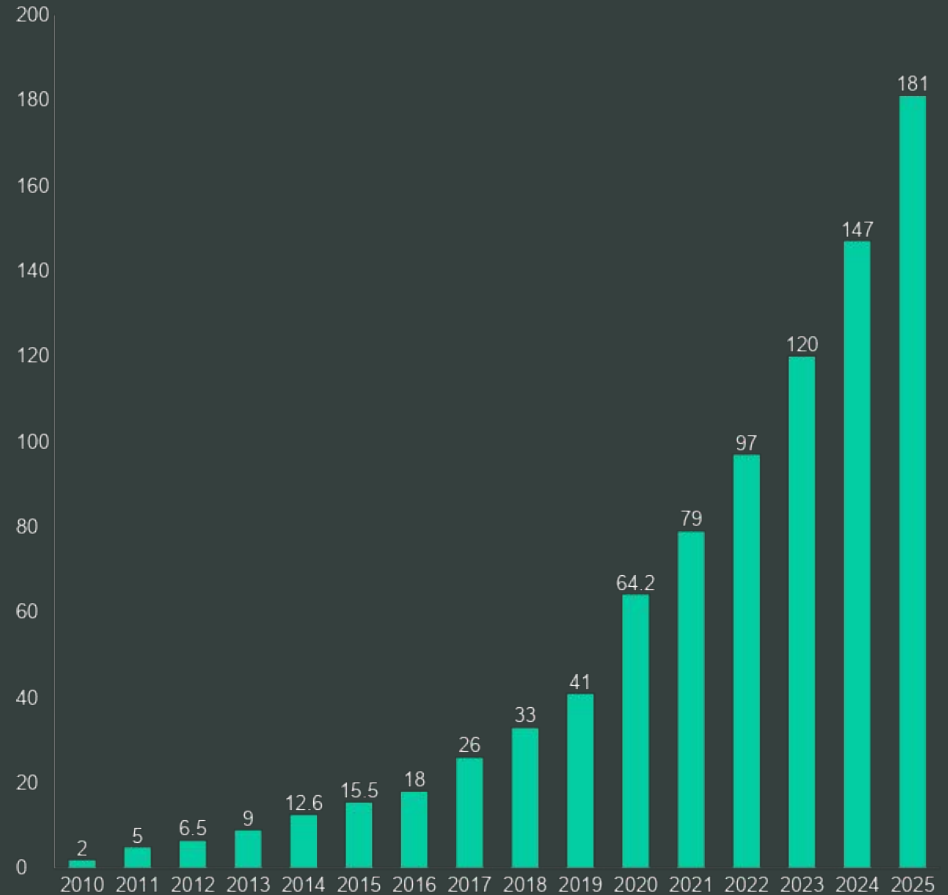


ACTIONABLE  
(USEFUL)



Data...

...the **problem**  
AND the opportunity





# Problem: The managed mainjor way to do it at a stack

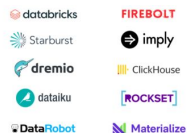
Choosing the best, voiced by customers

The vendor lock-in way

The open-source way

## Data 50 Companies: By Category

### QUERY & PROCESSING



### AI/ML



### ETL & ORCHESTRATION



### DATA GOVERNANCE & SECURITY



### CUSTOMER DATA ANALYTICS



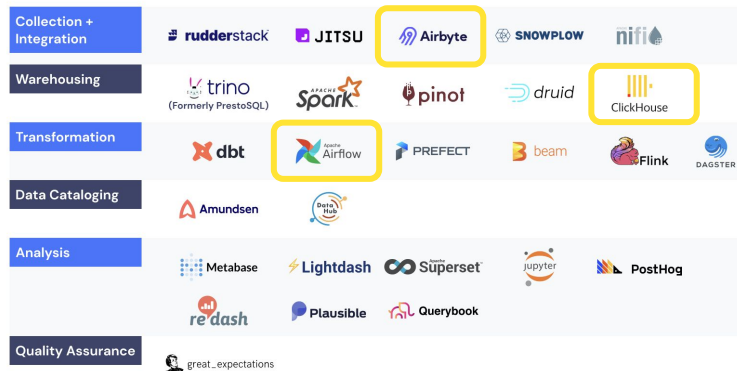
### BI & NOTEBOOKS



### DATA OBSERVABILITY



## The Modern Open-source Data Stack



# What is DoubleCloud?

DoubleCloud's platform helps you build sub-second data analytical solutions and pipelines on proven open-source technologies like ClickHouse® and Apache Kafka® in less than five minutes

1

## Cloud agnostic

Deploy on AWS, GCP and Azure (coming soon)

2

## Scalable

Use Clickhouse to build the fastest Data Warehouse for your analytics with up to 100+ shards or Apache Kafka to handle millions of events in real-time

3

## Low cost

Use hybrid storage to reduce the costs by 3x times. Compute based on ARM with 35% cost/value improvement. Free complimentary ETL and Visualization services

4

## Best in-class open-source as service

Build your analytics using open-source technologies with no vendor lock-in, unlike other proprietary services

5

## End-to-end analytics

Use tightly integrated services to build your observability stack or analyze millions of clicks and events in real-time with 3x less costs compared to other solutions



# DoubleCloud Platform

What technical value do we bring?



## Apache Kafka and Clickhouse

Automatic backups, patching, maintenance, cluster provisioning including sharding and replicas, and security, monitoring



## ETL and Visualization as SaaS

No servers to manage, auto-scale, free-tiers and native connectivity support to ClickHouse



## Free Backups and Free traffic

You don't need to pay for backups.  
Egress and cross-az traffic is free as well



## Free support 16/5

Free support for everyone, 16 hours per day, 5 days per week or upgrade for extended 24/7 support



## High Availability and Fault Tolerance

Resizing and configuration without downtime, high availability configurations automatically using different AZs without a single point of failure



## Security Out-of-the-Box

Encryption at rest out of the box on disk level and in transit. 2 levels of security groups from AWS and built-in



## Bring Your Own Account

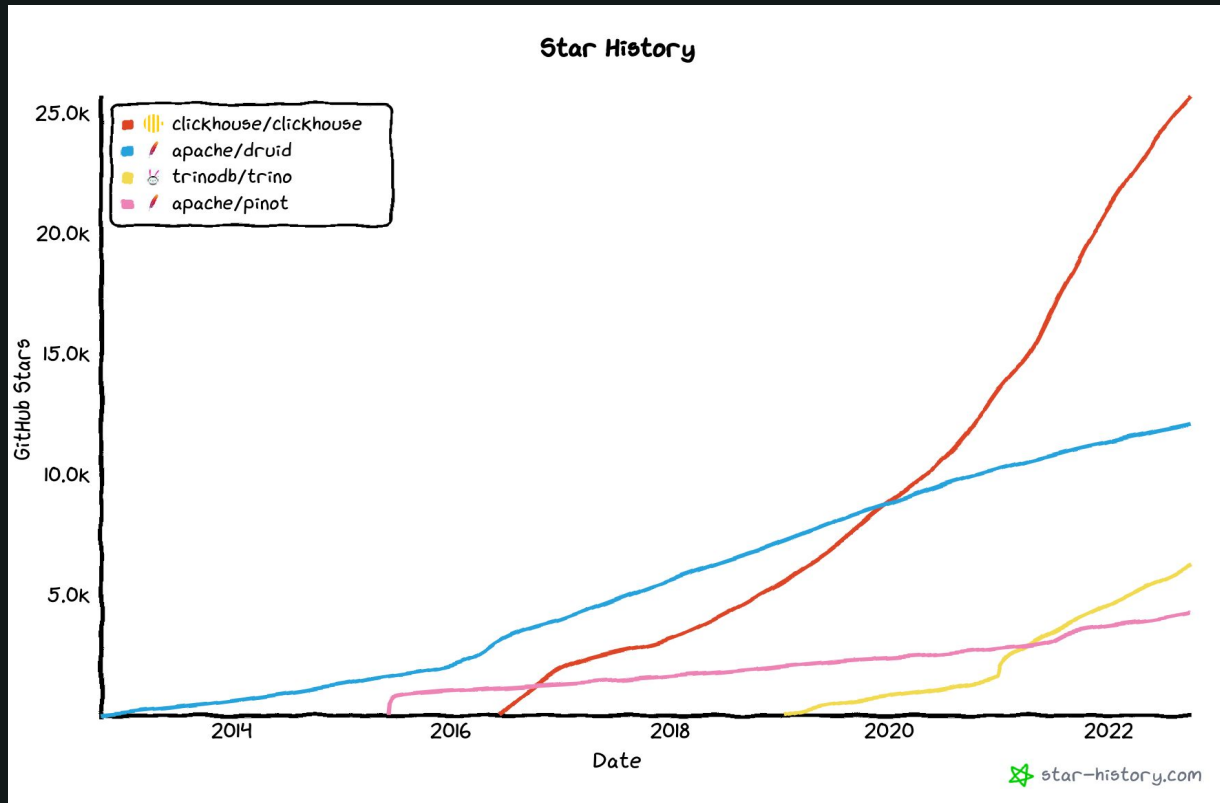
Ability to deploy clusters of Clickhouse and Kafka in your AWS account/VPC. All data and computation happen in your AWS account, and DoubleCloud manages the rest



## Hybrid Storage - cost effective

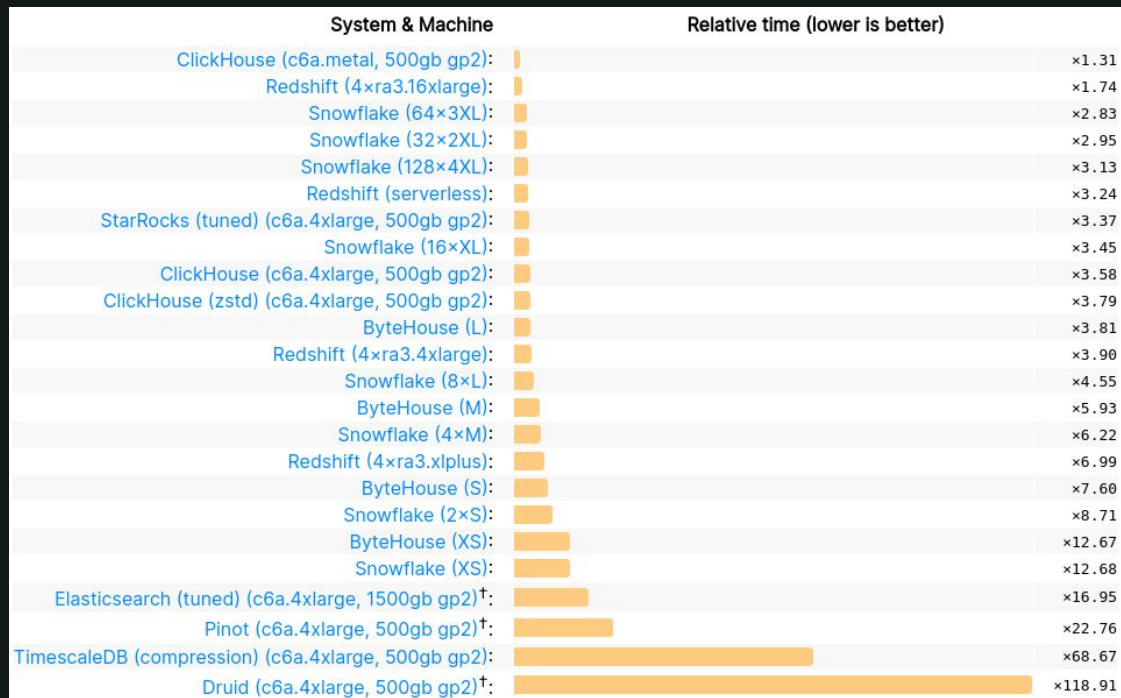
Store data on SSD and S3 transparently. Automatically decoupling the latest or most frequent data to SSD and the less frequent data to S3, reducing the cost for storage up to 3x times

# What is ClickHouse?



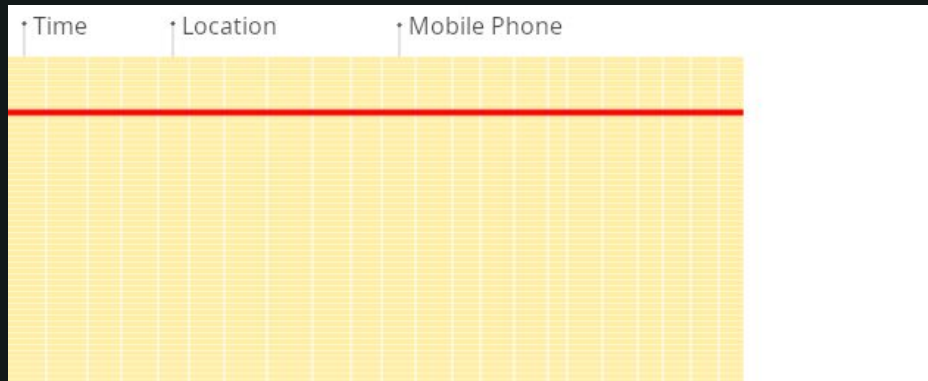
See: <https://github.com/clickhouse/clickhouse>

# What is ClickHouse?



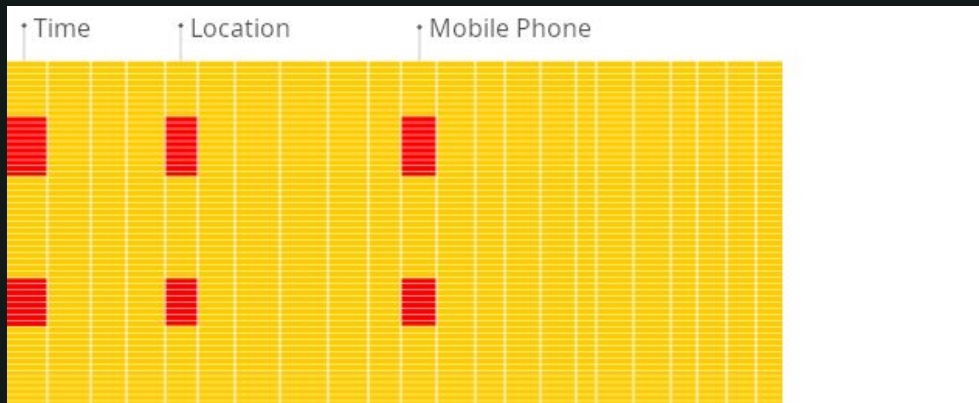
See: <https://benchmark.clickhouse.com/>

# Columnar vs row



Time	Location	Mobile Phone
------	----------	--------------

Row by row –  
Great for inserts and row updates

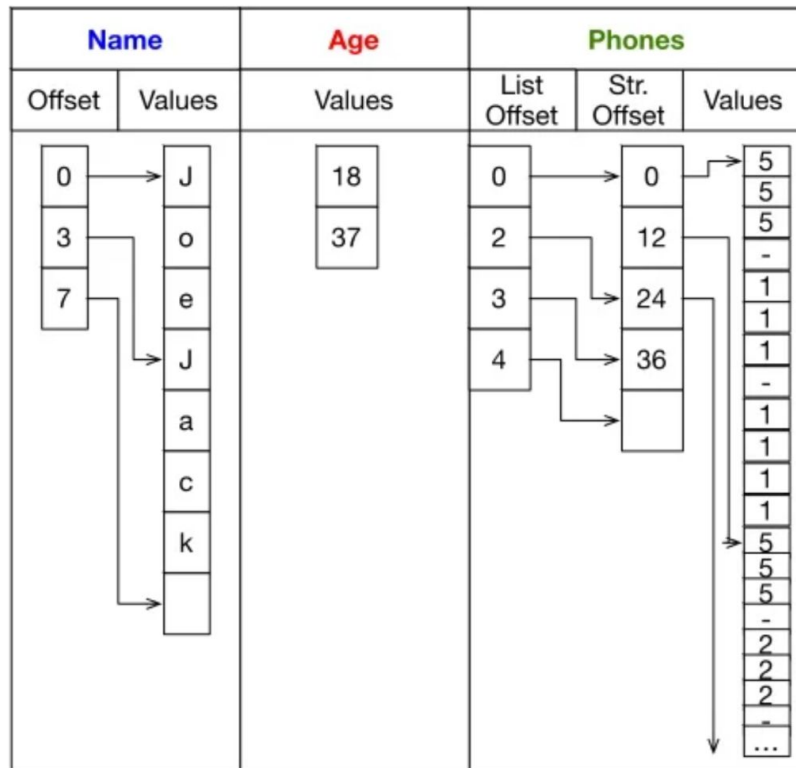


Time	Location	Mobile Phone
------	----------	--------------

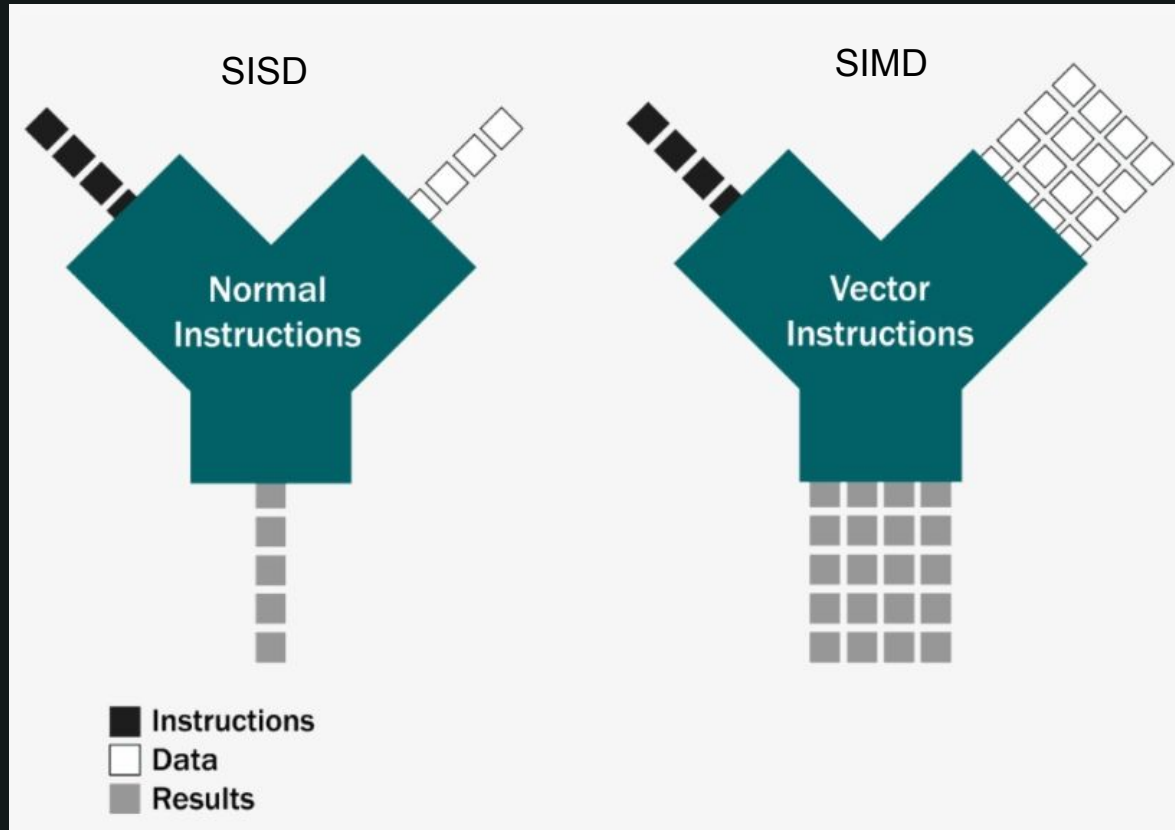
Columnar –  
Great for large bulk operations

## Columnar data

```
persons = [{
    name: 'Joe',
    age: 18,
    phones: [
        '555-111-1111',
        '555-222-2222'
    ]
}, {
    name: 'Jack',
    age: 37,
    phones: [ '555-333-3333' ]
}]
```



# Vector execution - SUPER FAST!





# ClickHouse Adopters



Uber



Yandex

Self-Driving  
Group

Practicum

PicsArt



# Ebay - case story



**“we reduced our overall infrastructure footprint by over 90”**

**“10 times less hardware”**

**“Stronger integration with Grafana”**

eBay adopted ClickHouse for their real time OLAP events (Logs + Metrics) infrastructure. The simplified architecture with ClickHouse allowed them to reduce their DevOps activity and troubleshooting, reduced the overall infrastructure by 90%, and they saw a stronger integration with Grafana and ClickHouse for visualization and alerting.

# Uber - case story

## Uber

“Writes 3x - 4x throughput compared to ES”

“Ingest performance scales close to linearly to cluster size”

“Multi-master replication ensures no SPOF in design”

“~5x query speed of ES”

# Spotify - case story



“Spotify’s A/B Experimentation platform is serving **thousands of sub-second queries per second** on **petabyte-scale** datasets with Clickhouse.

They reduced the amount of low-variance work by an order of magnitude and **enabled feature teams to self-serve insights** by introducing a unified SQL interface for Data Platform and tools for **automatic decision making** for Experimentation”

# Cloudflare - case story



**“Scaled to 8M requests per second”**

**“Reduced their MTTR (Mean time to repair)”**

**“7x improvement on query throughput”**

“Cloudflare was having challenges scaling their CitusDB-based system which had a high TCO and maintenance costs due to the complex architecture. By moving their HTTP analytics data to ClickHouse they were able to scale to 8M requests per second, deleted 10's of thousands of lines of code, reduced their MTTR, and saw a 7x improvement on customer queries per second they could serve.”

# Companies using ClickHouse with DoubleCloud

Yandex

Self-Driving  
Group

Practicum

PicsArt

 *Honeybadger*

 Toloka



QIP

 *hystax*

 THE  
UNIVERSITY  
OF UTAH





AN annisudoublecloud ▾

Clusters



## Clusters

[Create cluster](#)

NAME

STATUS

REGION

CREATED

No data



Create a Clickhouse cluster

# Interfaces

- HTTP
- Native TCP
- gRPC

**Support native wire protocols for:**

- MySQL
- PostgreSQL

## Official clients



## Language client libraries



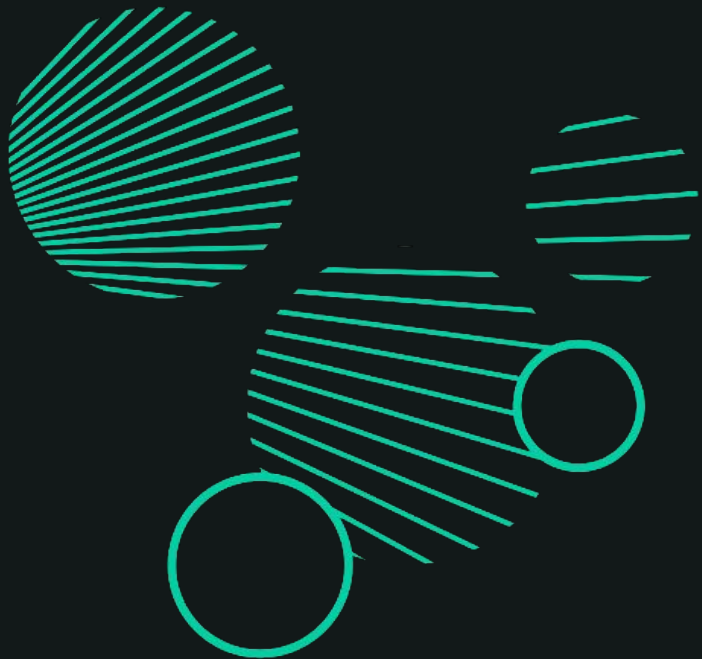
# File Formats

## Nearly Every File Format

- Avro
- Arrow
- CSV
- URL
- Prometheus
- ORC
- Parquet
- Hudi
- deltaLake
- Protobuf
- XML
- JSON




Tool	Stars (GitHub)	Processing Time (Map/Aggr/Filter)	Memory Scalability (Map/Aggr/Filter)	Conclusion
<a href="#">ClickHouse</a>	26k	👍👍👍	👍👍👍	Overall the fastest for large files (>=100MB).
<a href="#">OctoSQL</a>	4.2k	👍👍👍	👍👍👍	Overall the fastest for small files (1-10MB), head to head with ClickHouse on larger files.
<a href="#">SPyQL</a>	842	👍👍👍	👍👍👍	Up to 2x faster than jq but up to 5x slower than the best (for 1GB of data). 2nd lowest memory footprint (22MB), independently of the input size.
<a href="#">jq</a>	24k	reference	👍👍👍	The lowest memory footprint (6MB) if you can avoid building arrays in memory.
<a href="#">Miller</a>	5.8k	👍👍👎	👎👎👎	Comparable or faster performance than jq. Memory grows with the size of input data => always reads the full dataset into memory.
<a href="#">trdsq!</a>	1.3k	👎👎👎	👍👍👍	Among the slowest but always with a low memory footprint (29MB). Always processes the full dataset.
<a href="#">spark-sql CLI</a>	34k	👎👎👎	👎👍👍	Large overhead, catching-up on GB-sized data, but always among the slowest. Always processes the full dataset. Output size does not scale (full output is written to the driver's memory in the CLI's implementation).
<a href="#">DSQ</a>	2.9k	👎👎👎	👎👎👎	Among the slowest and with a large footprint => always reads the full dataset into memory.

# Managed Kafka



## Clusters





[Create cluster](#)

NAME	STATUS	REGION	CREATED	
 c32m128hd512 Elasticsearch 7.10.2.8	<span>Stopped</span>	eu-central-1 AWS	08/28/2022 14:25 UTC	...
 ch-c4m16hd384-w-kafka Elasticsearch 7.10.2.8	<span>Stopped</span>	eu-central-1 AWS	09/08/2022 17:45 UTC	...
 kafka-c4-m16 Kafka 3.0.1	<span>Stopped</span>	eu-central-1 AWS	09/14/2022 16:17 UTC	...

# Managed Kafka Interface

## Clusters

[Create cluster](#)

NAME	STATUS	REGION	CREATED	
 c32m128hd512 ClickHouse · v22.9	<span>Active</span>	eu-central-1 AWS	08/26/2022 14:25 UTC	...
 ch-c4m16hd384-w-kafka ClickHouse · v22.9	<span>Stopped</span>	eu-central-1 AWS	09/08/2022 17:45 UTC	...
 kafka-c4-m16 Kafka · v3.1	<span>Active</span>	eu-central-1 AWS	09/14/2022 16:17 UTC	...
 kafka-cluster Kafka · v3.1	<span>Creating</span>	us-east-1 AWS	10/05/2022 02:06 UTC	...

# Transfer Service

- Marketing/Ads:



- Databases:



- Warehouses:



- Others:



And more...

Manage Source Endpoints in Transfer  
Create a source endpoint



Note

The Transfer service uses [Airbyte](#) to migrate data.



# DoubleCloud Transfer

A cloud agnostic service for aggregating, collecting, and migrating data from various sources

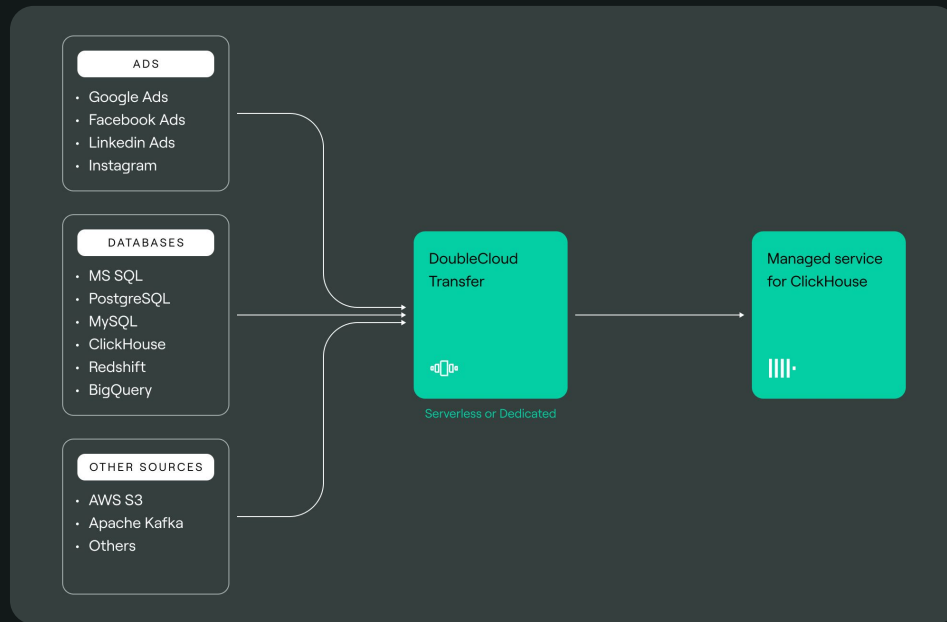
Aggregate and load data from different sources to build unified analytics

27

Offload analytical workloads from transactional databases — e.g. from PostgreSQL

Establish real-time replication across databases or even different type of databases

Easily migrate data from or to on-premise or other places



# DoubleClick Visualization

Free cloud-based BI tool. Quickly test hypothesis and do ad-hoc analytics

+15 different type of charts, including maps and heatmaps

Ability to connect to ClickHouse, MSSQL, PostgreSQL, MySQL, csv on S3

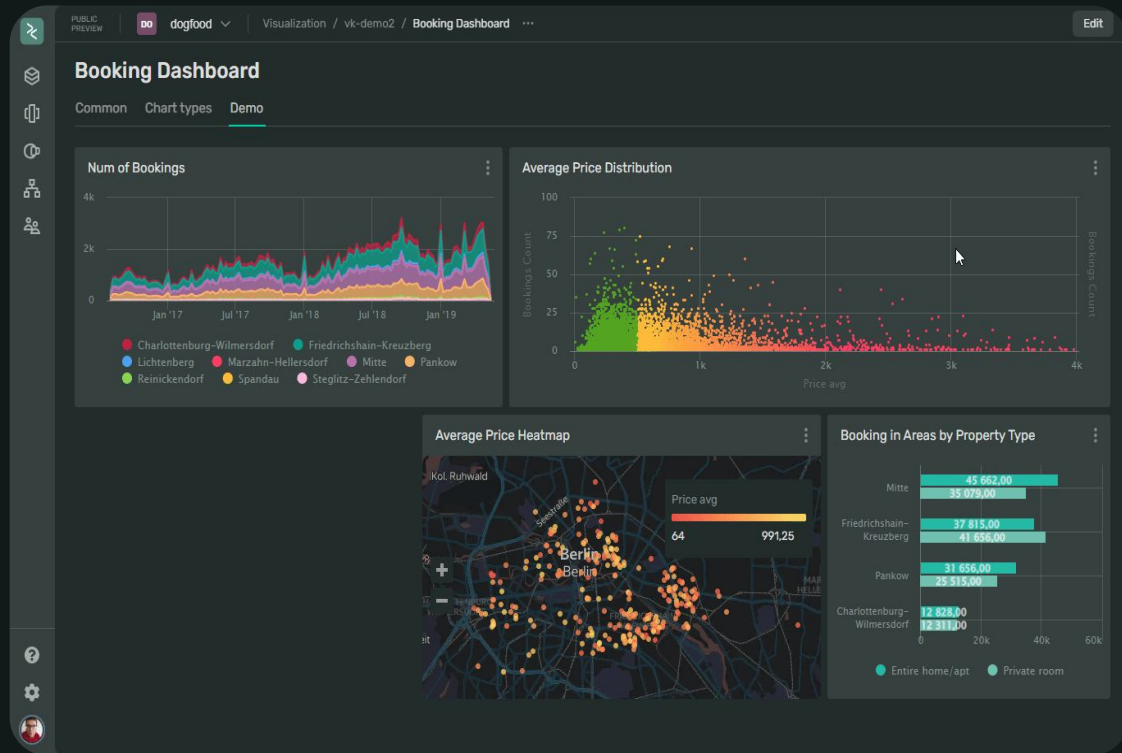
Self BI with Web UI interface service

Row-based security and filtering

Custom SQL queries, public dashboards, parametrized charts with drill downs and many more other features

**Totally Free of charge**

28



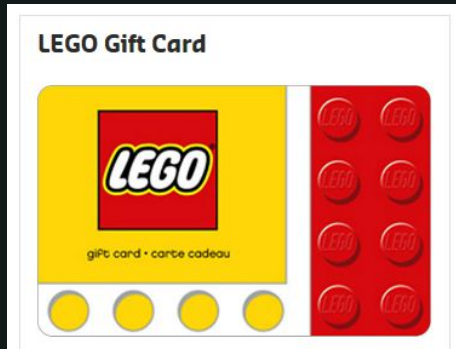
# Q&A and Call to

• Questions?

• Visit Our Booth

• Book a Demo

• Interested in Working at DoubleCloud?





Follow us **@goto\_con**  
Share your experience using **#YOWlondon**

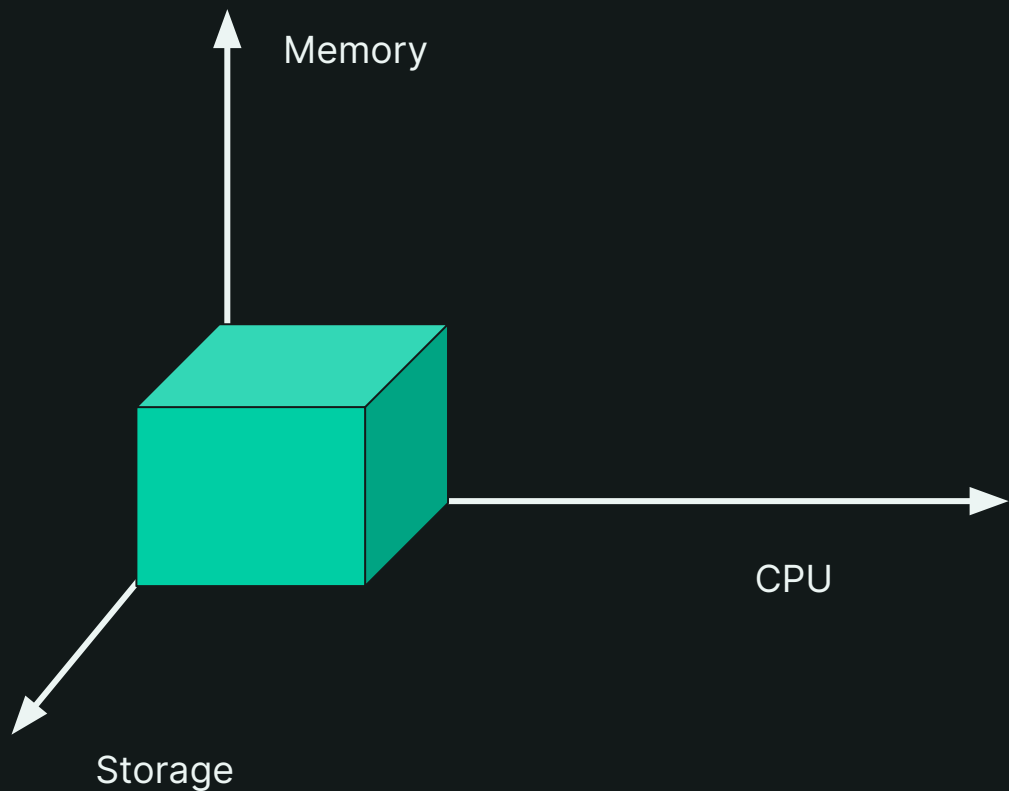


Follow us **@GOTOcon**  
Share your experience using **#YOWlondon**

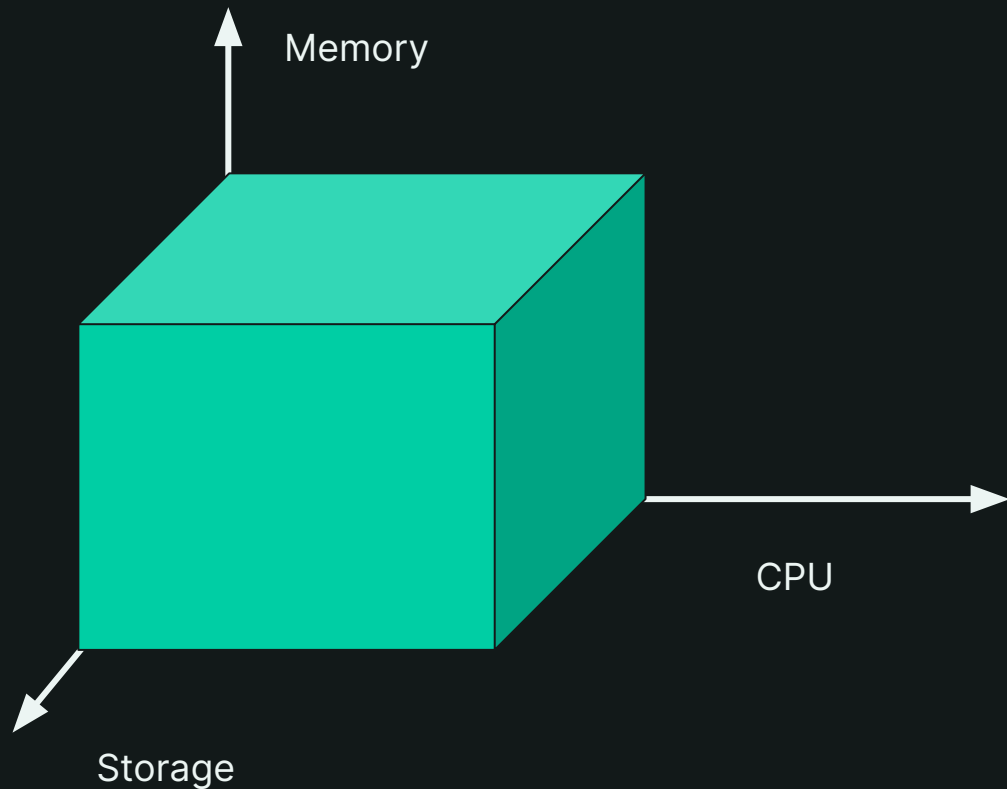


Follow us **GOTO Conferences**  
Join **322k+ subscribers & 36m+ views**

ClickHouse  
nodes  
can scale  
vertically

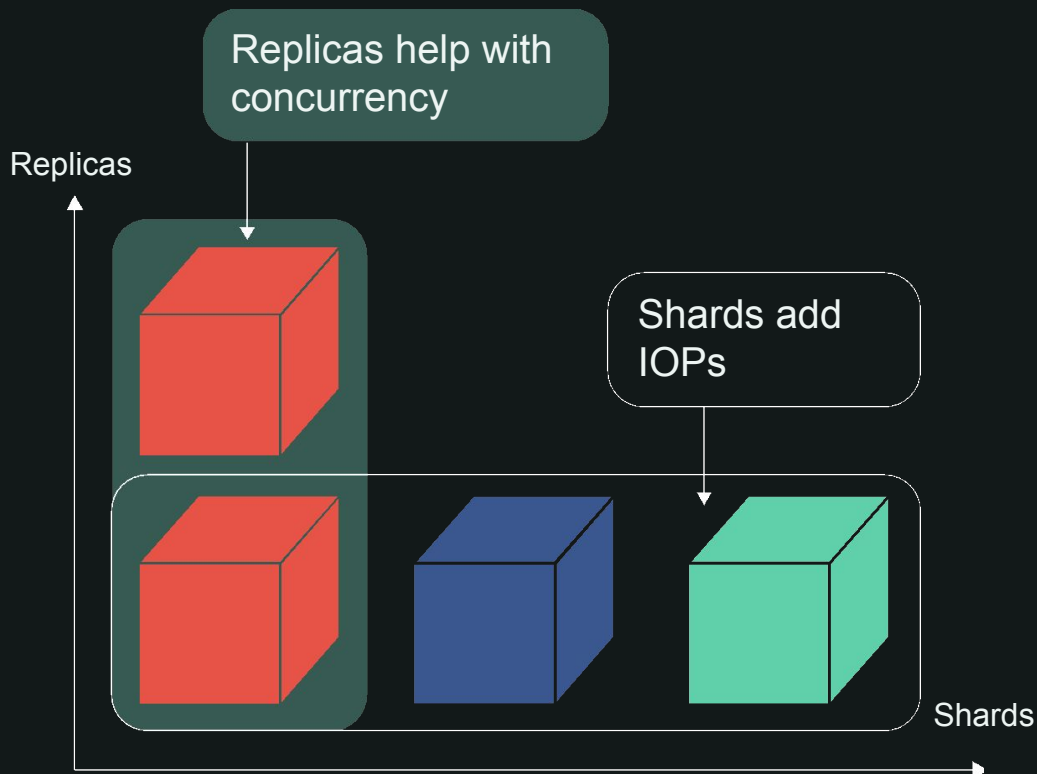


ClickHouse  
nodes  
can scale  
vertically





ClickHouse  
nodes  
can scale  
horizontally



# Different sharding and replication patterns



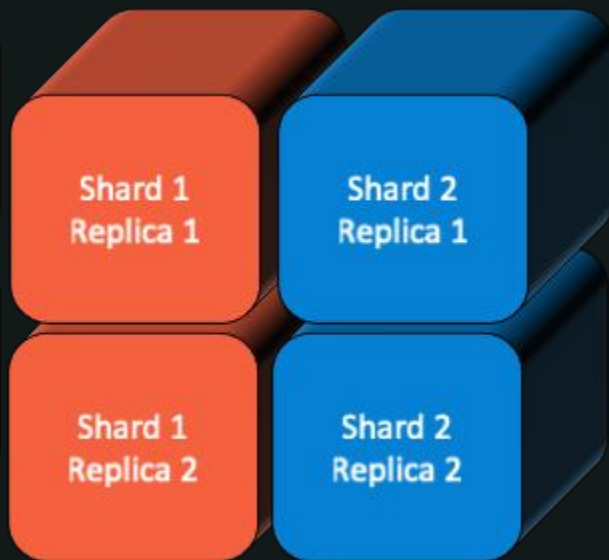
## All sharded

Data sharded 4 ways without replication.



## All replicated

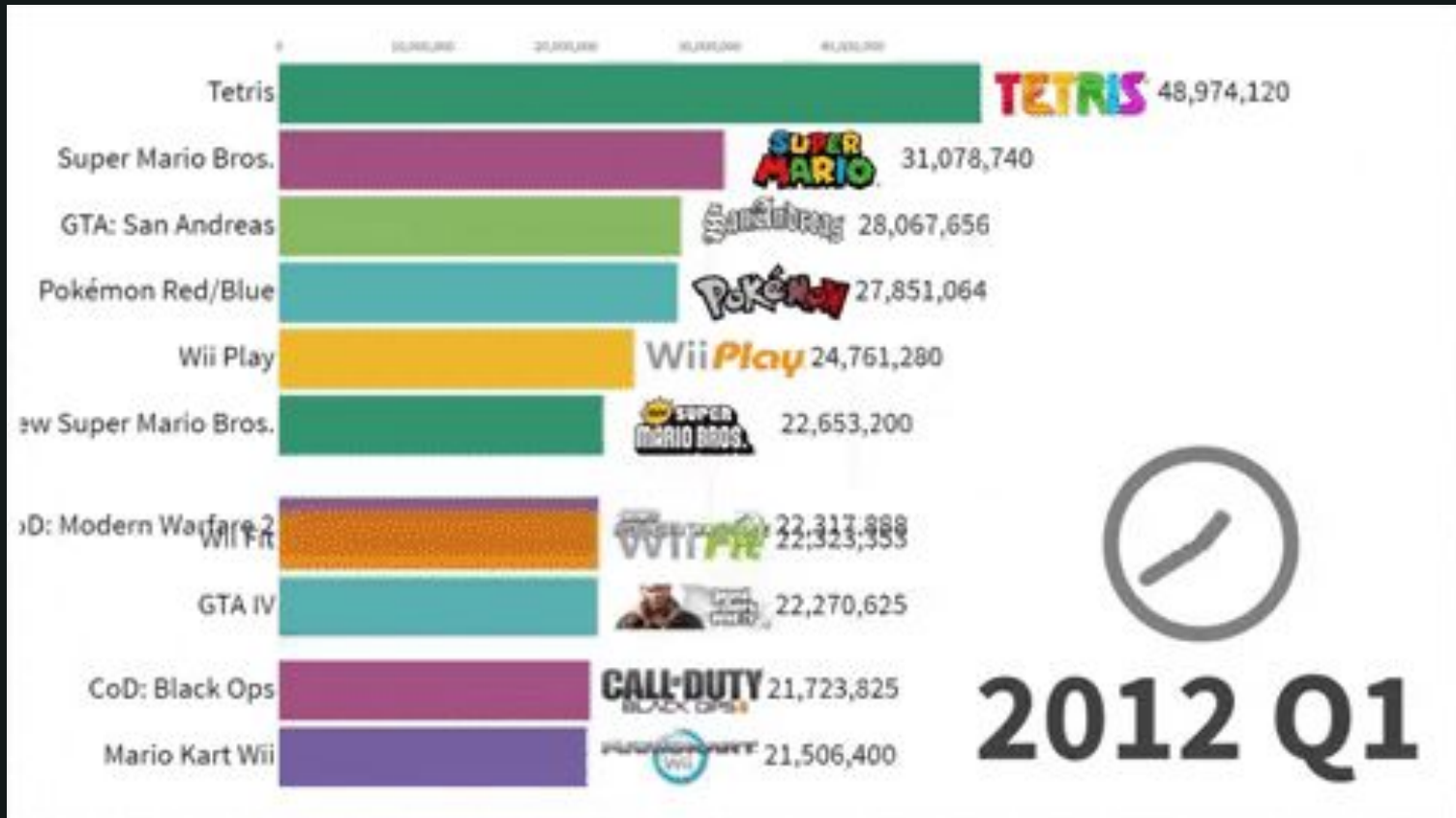
Data replicated 4 times without sharding.



## Sharded and Replicated

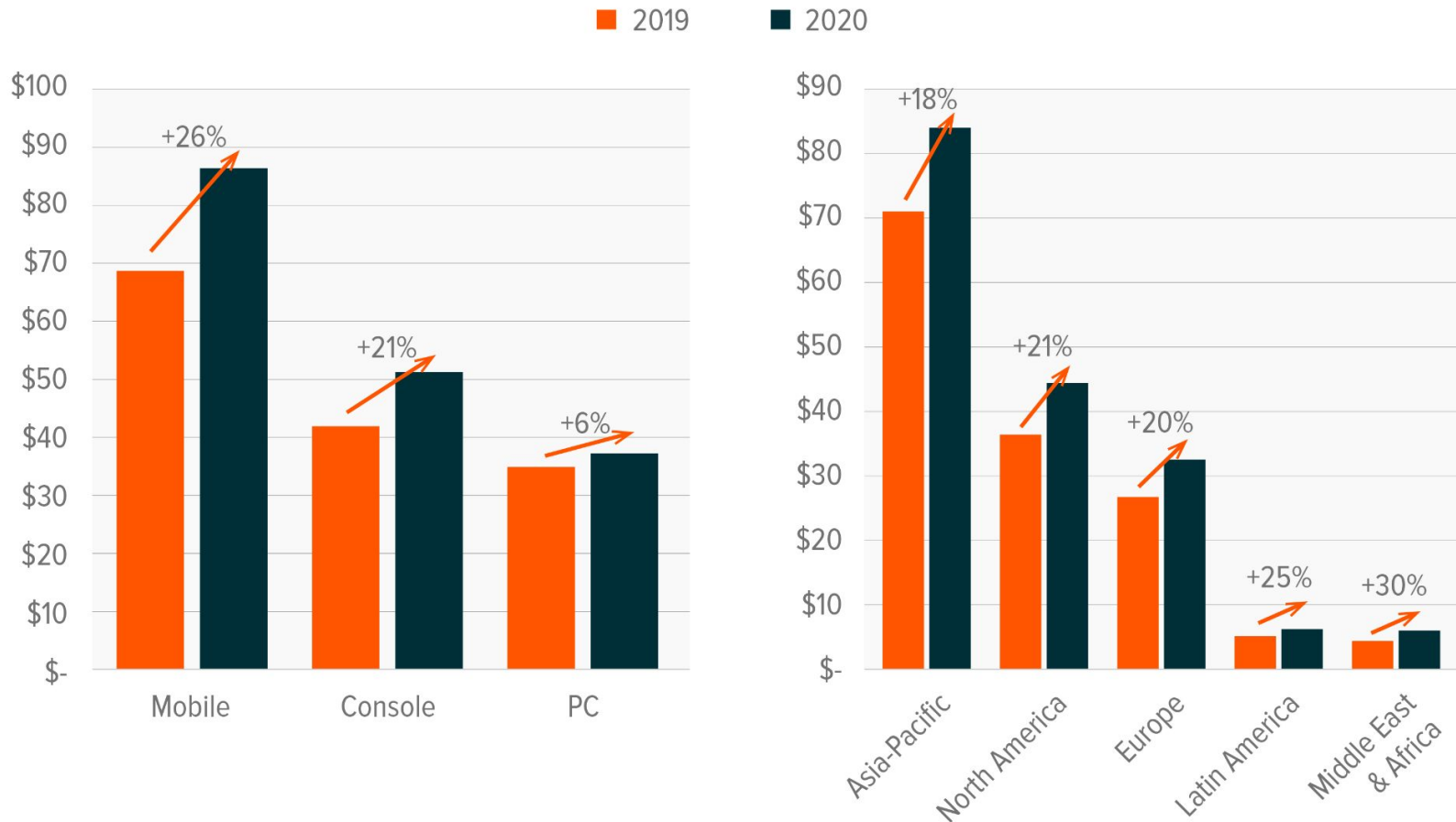
Data sharded 2 ways and replicated 2 times.

# The rise of MINECRAFT with data visualization



# VIDEO GAMES REVENUES: BY DEVICE AND GEOGRAPHY (IN BILLIONS)

Source: Newzoo. Estimates as of Nov 2020.





DATA

#1



SORTED

#2



ARRANGED

#3



PRESENTED  
VISUALLY

#4



EXPLAINED  
WITH A STORY

#5



ACTIONABLE  
(USEFUL)

#6

# DoubleCloud Transfer Features:

CDC and ETL is **easy with no-code**

17+ Native connectors: PostgreSQL, 38  
MySQL, S3, BigQuery, Redshift and many  
others from the community

Auto-scaling upto 1gb/sec in serverless  
mode

Dedicated mode for predictable  
performance in secure connectivity

Supports VPC peering and BOYA

Based on **open-source** and **Airbyte**  
connectors

Observability of transfer process: audit,  
monitoring, alerts, progress of snapshotting

Simple transformations

Data service name	Snapshot	Increment	Snapshot and Increment
ClickHouse®	<input checked="" type="checkbox"/>		
Apache Kafka®		<input checked="" type="checkbox"/>	
MySQL			<input checked="" type="checkbox"/>
Amazon Redshift	<input checked="" type="checkbox"/>		
Google Ads	<input checked="" type="checkbox"/>		
Facebook Marketing		<input checked="" type="checkbox"/>	
LinkedIn Ads	<input checked="" type="checkbox"/>		
AWS CloudTrail	<input checked="" type="checkbox"/>		
MS SQL	<input checked="" type="checkbox"/>		
Amazon Ads	<input checked="" type="checkbox"/>		
Snapchat Marketing	<input checked="" type="checkbox"/>		
Instagram	<input checked="" type="checkbox"/>		
PostgreSQL			<input checked="" type="checkbox"/>
MongoDB			<input checked="" type="checkbox"/>
Amazon S3	<input checked="" type="checkbox"/>		
BigQuery	<input checked="" type="checkbox"/>		



# DoubleCloud Visualization

## Scenarios and Use-Cases

Quickly test hypotheses using real data

Gather key business metrics from different sources into a single dashboard

Share resulting analytics with your team

Ad hoc data analysis and visualization including:

- Web analytics
- Mobile app analytics
- Machine learning result analysis
- Partner analytics
- Geoanalytics
- Open data and public analytics

