Extreme Programming *meets* Real Time Data

Gel Goldsby & Tom Johnson, Unruly

When Santa Got Stuck Up The Chimney

When Data Got Stuck Up The Chimney

Hello. My name is...

Gel Goldsby
Reporting and
Data Team Lead

Tom Johnson
Senior Developer

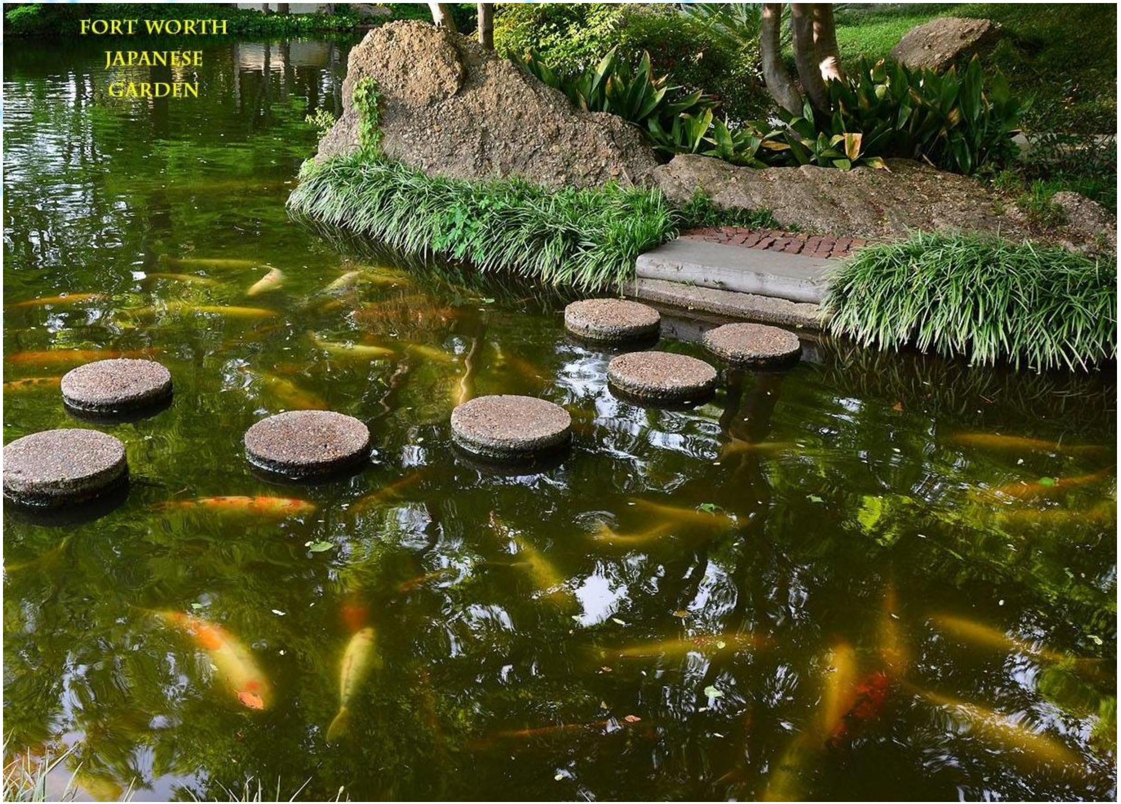
UNRULY

We Believe In XP

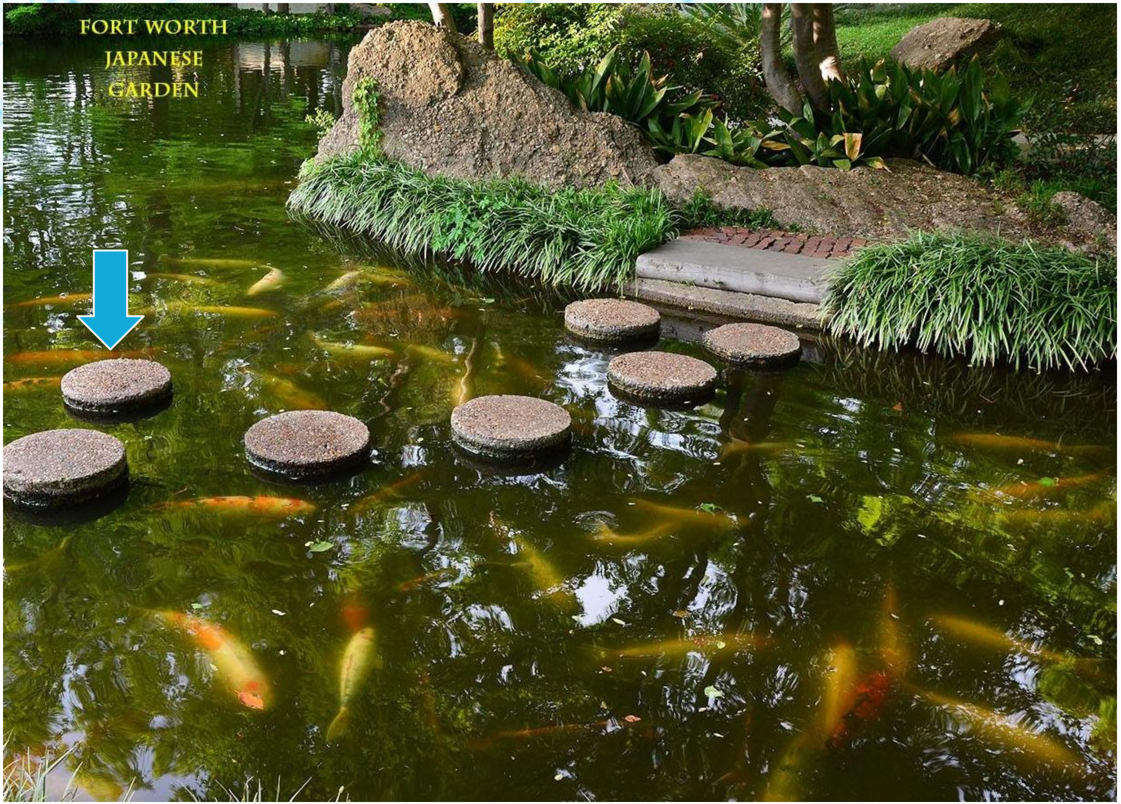
Extreme Programming Values

- Communication
- Simplicity
- Feedback
- Courage

Simplicity



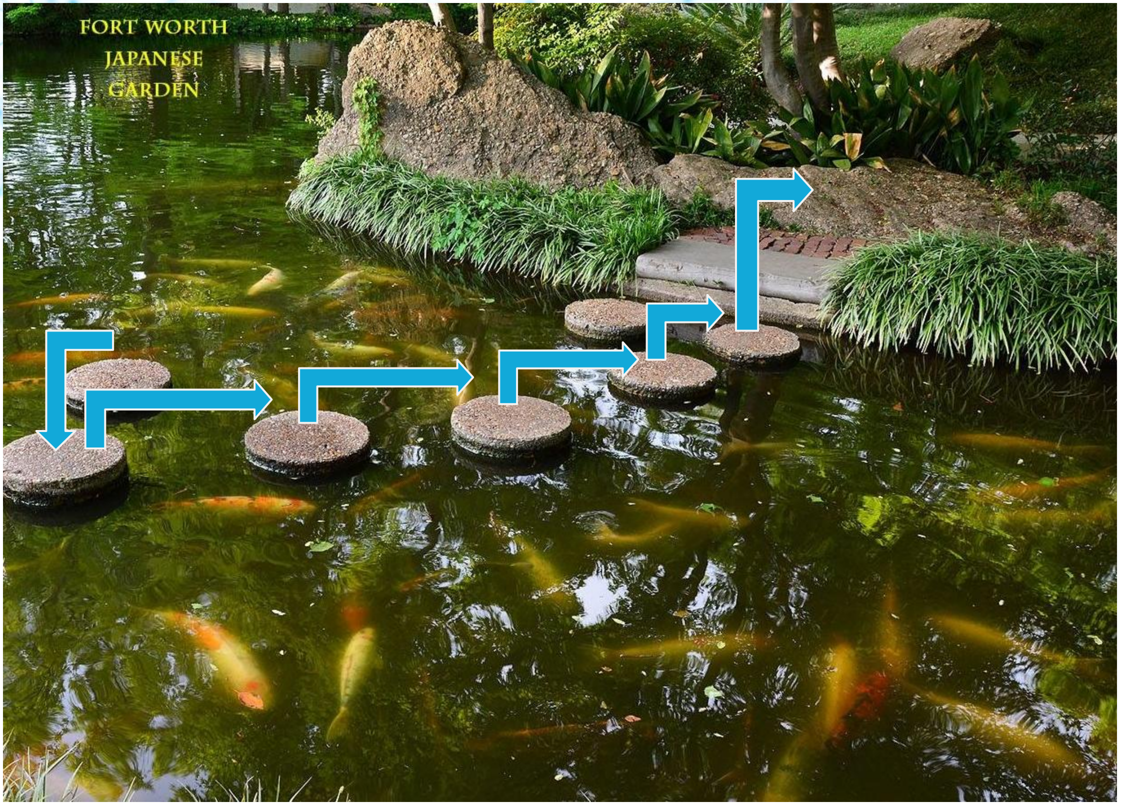
Simplicity



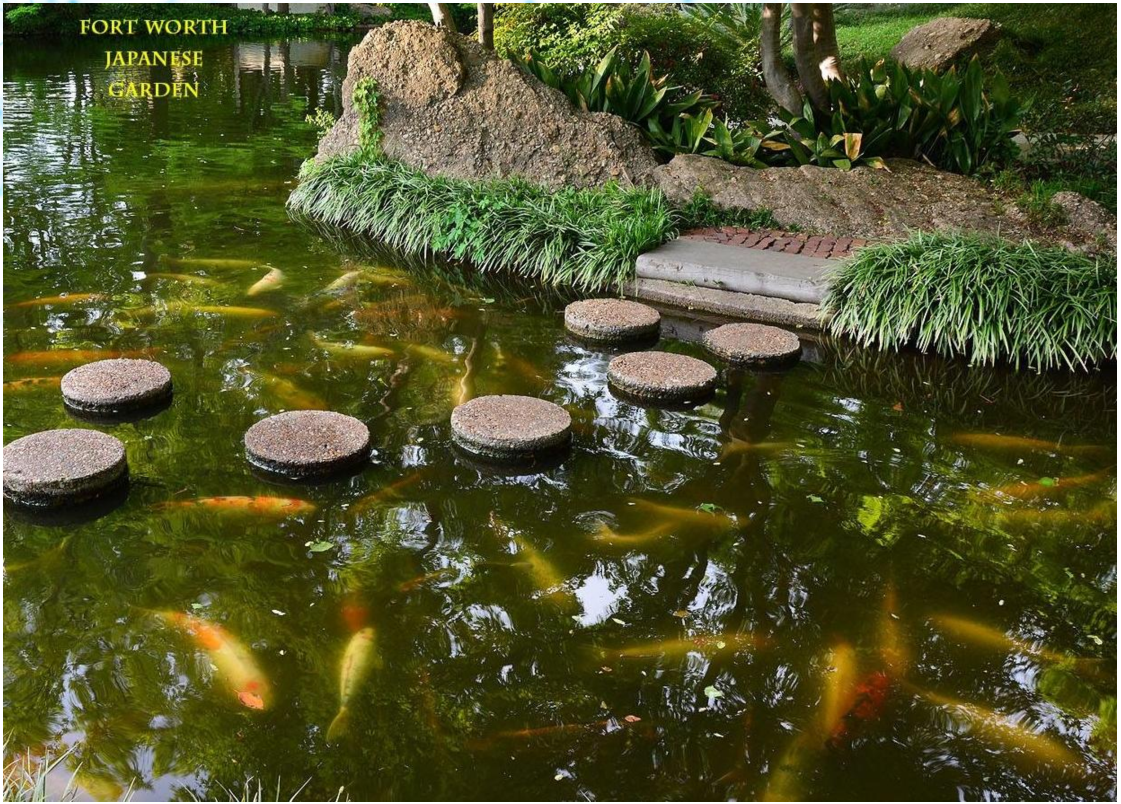
Simplicity

Simplicity

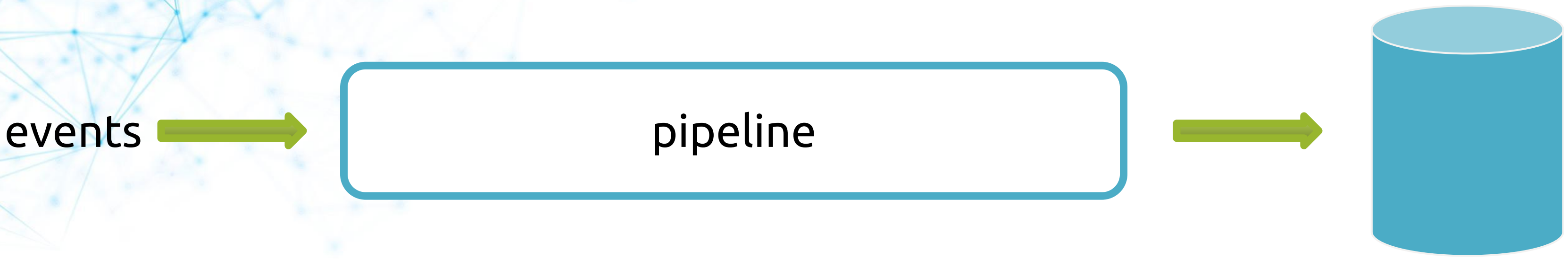
Simplicity



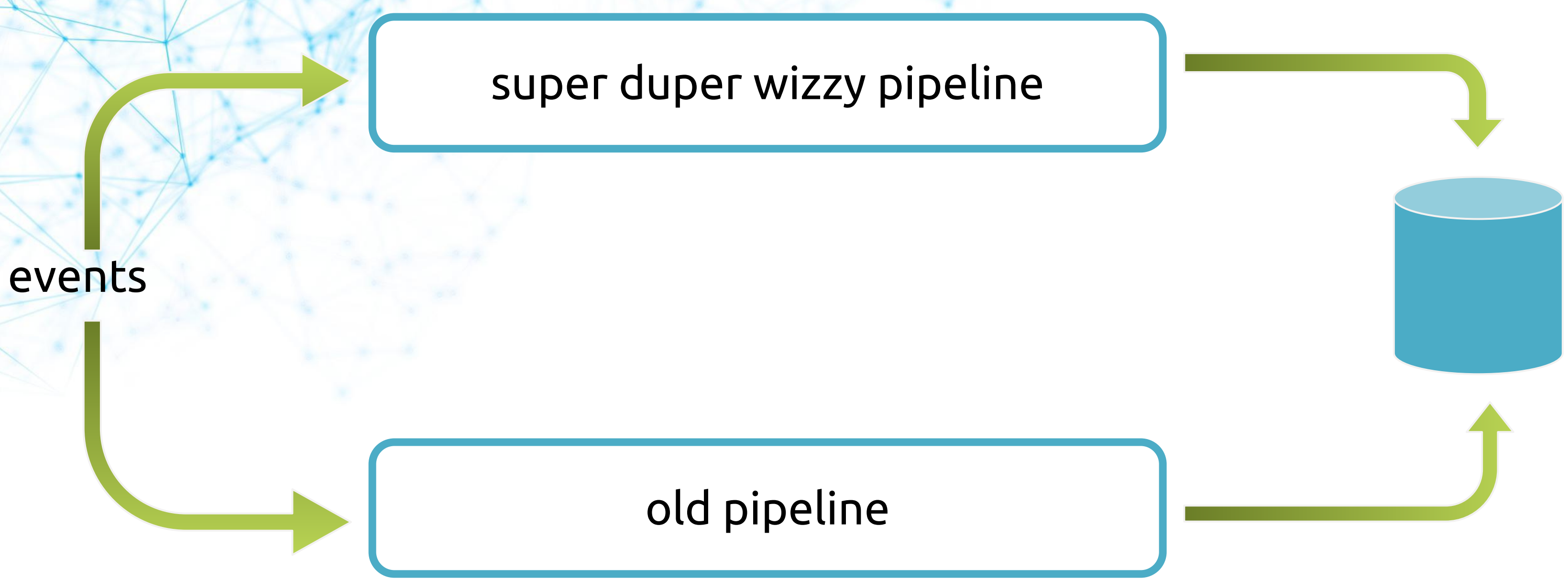
Simplicity



Our Reporting Pipeline



Our Reporting Pipelines

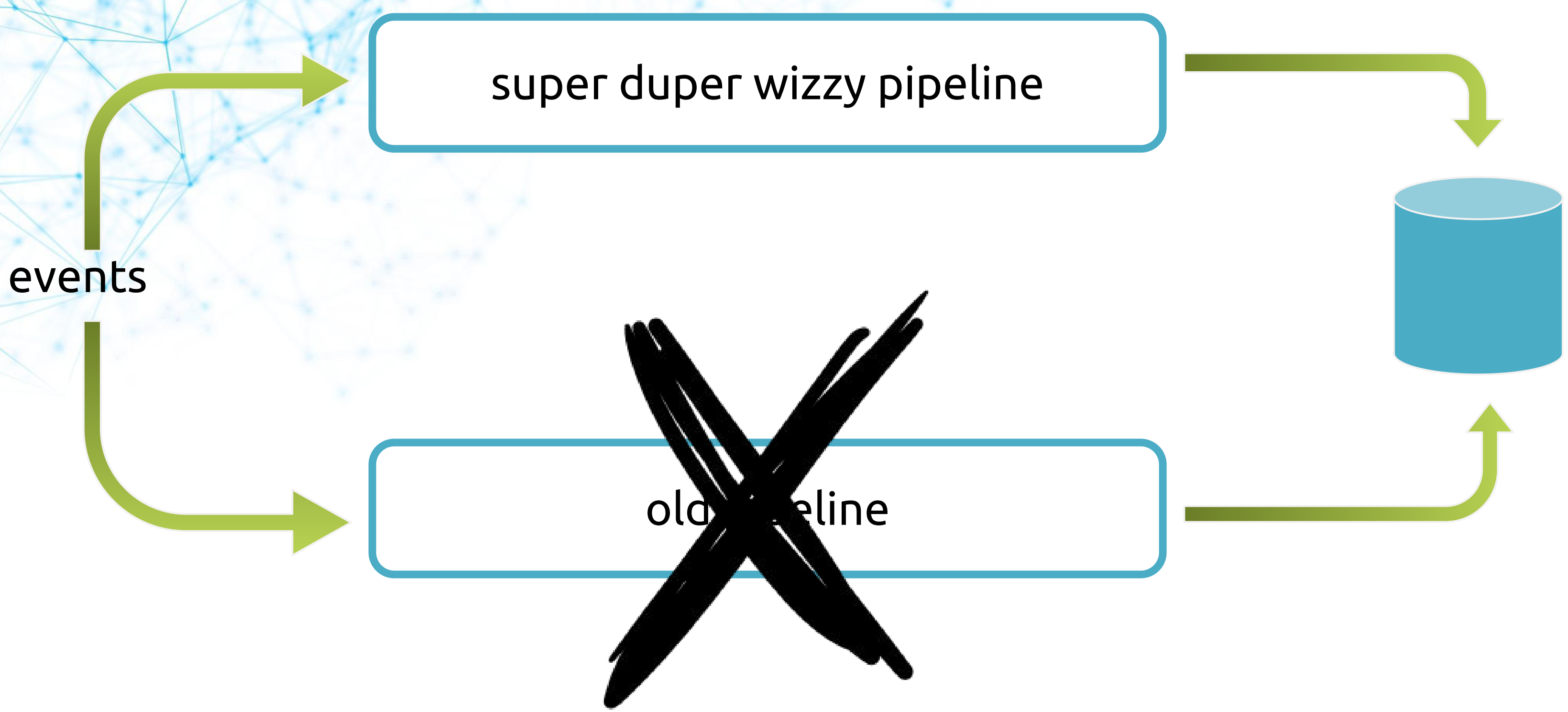


events

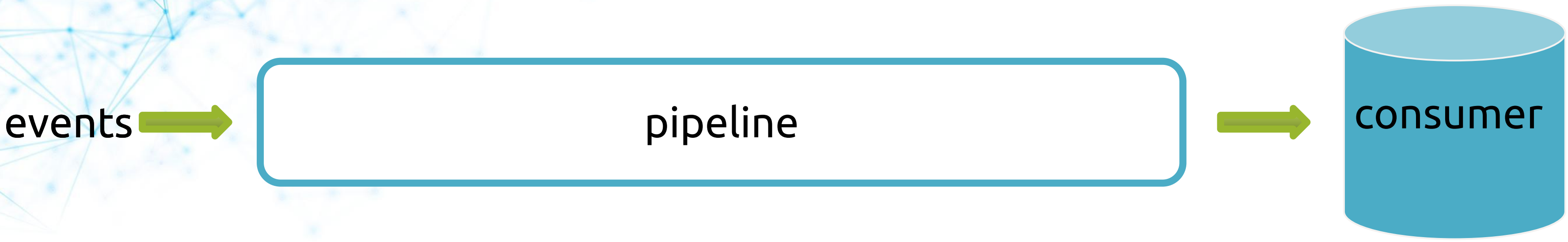
super duper wizzy pipeline

old pipeline

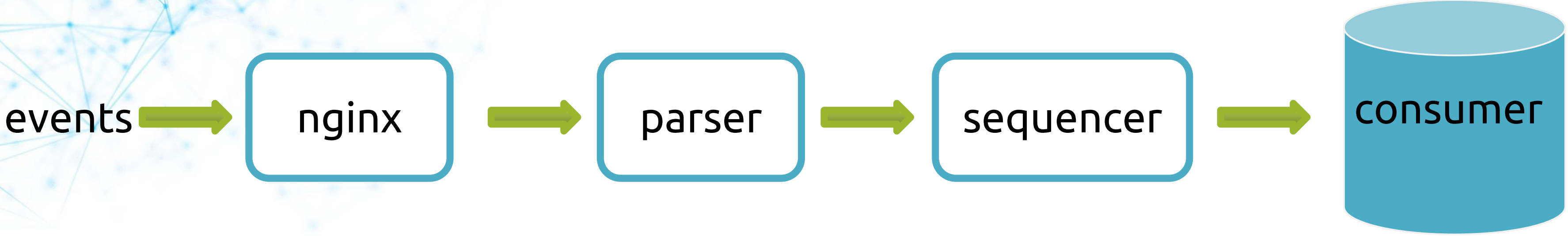
Shut It Off!



A Closer Look At Our Pipeline

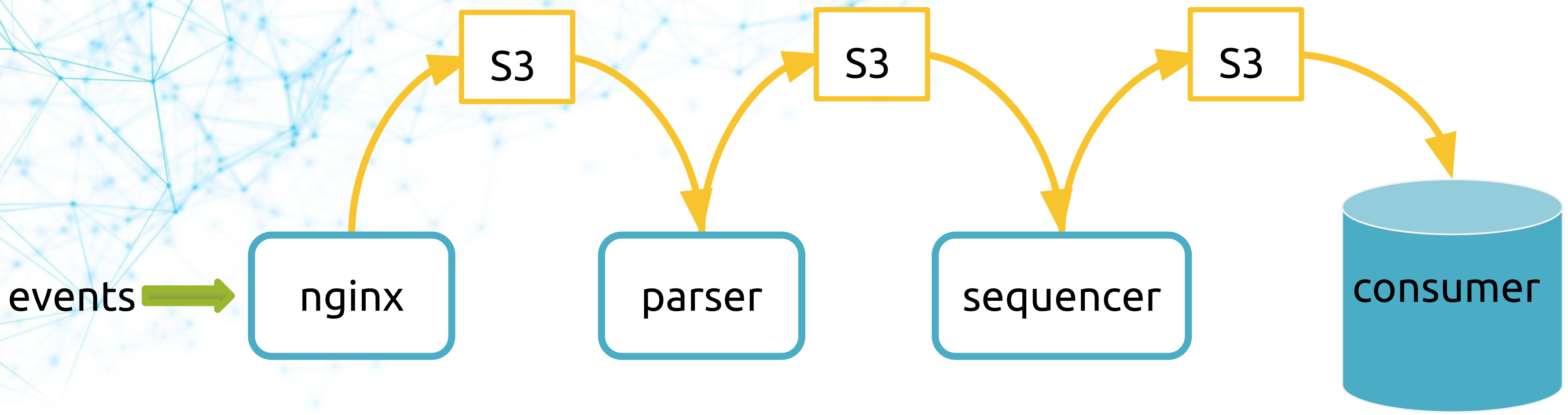


It’s Not A Truck, It’s A Series of Tubes

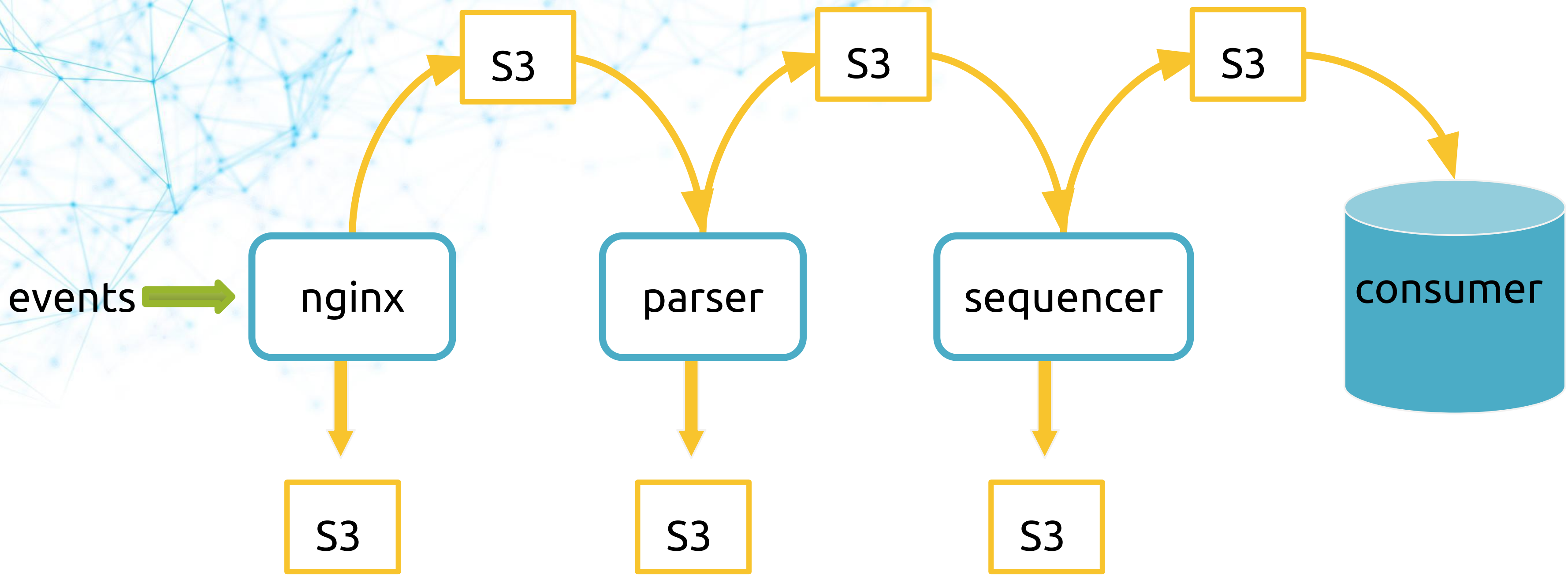


events → nginx → parser → sequencer → consumer

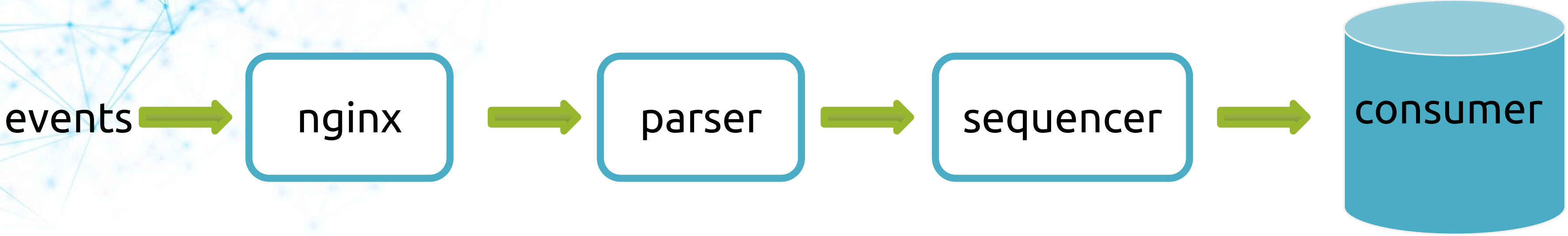
Queueing with S3



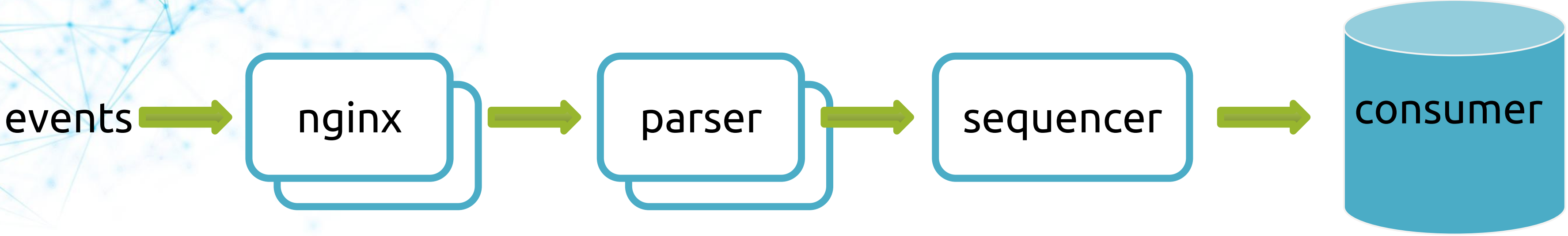
Queueing with S3



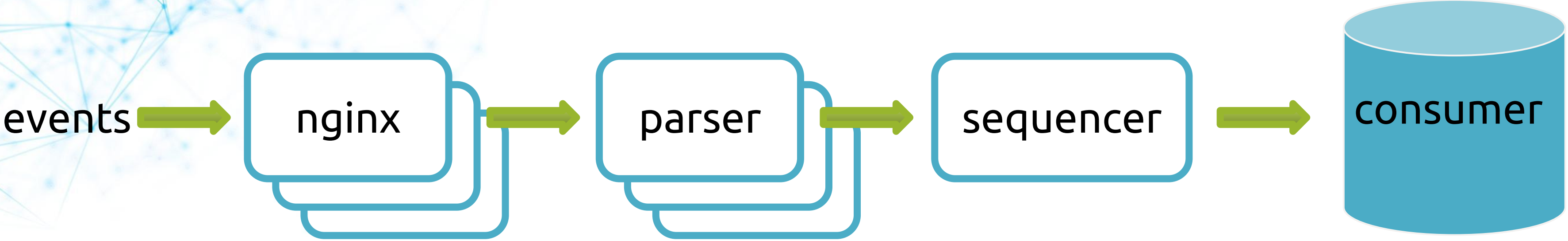
We Need More Power, Cap’n



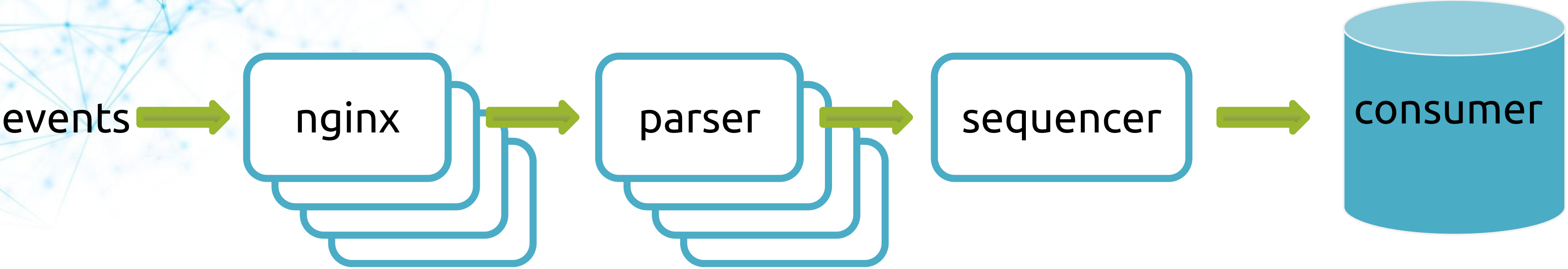
We Need More Power, Cap’n



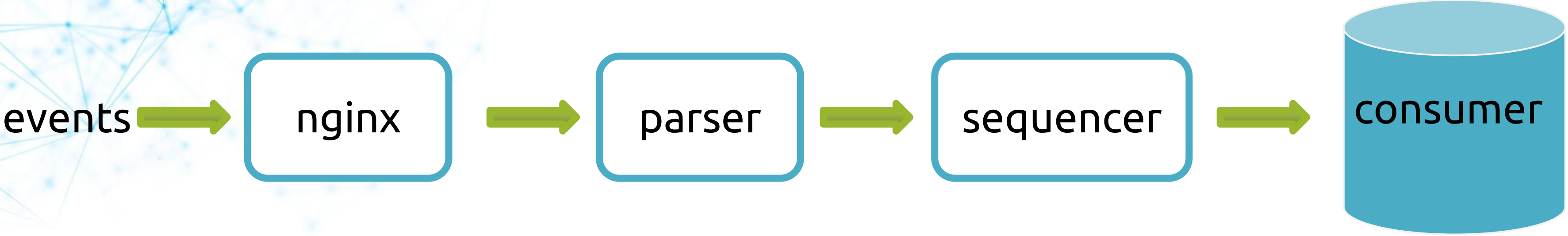
We Need More Power, Cap’n



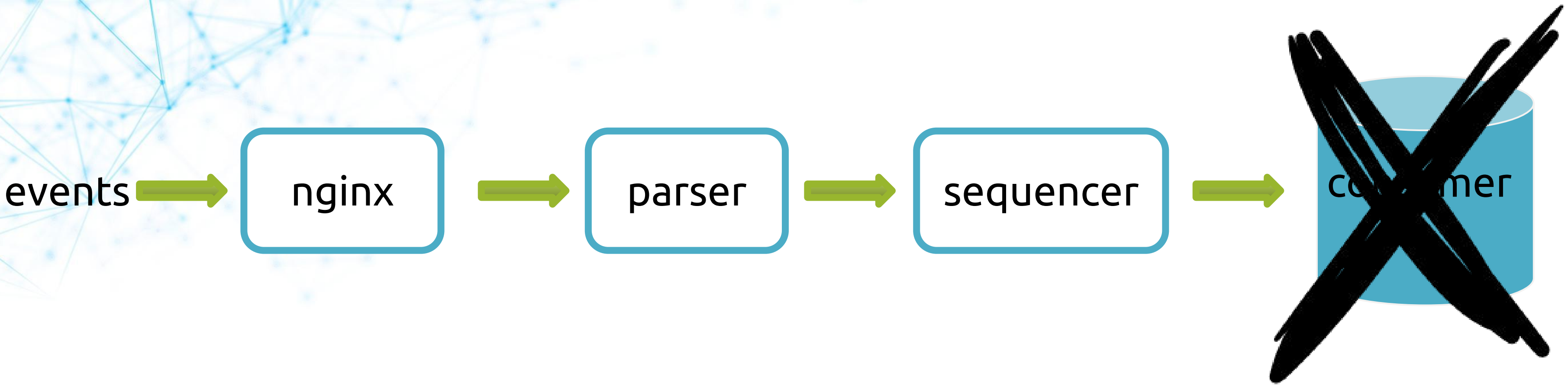
We Need More Power, Cap’n



Two Writes Can Make A Wrong



Two Writes Can Make A Wrong



events

nginx

parser

sequencer

~~consumer~~

Christmas was saved!

Simplicity

- Each component does one thing and does it well

Just Another Report, Right?

- Improving targeting
- Correlate events for same ad call
- Need to join on session id
- Needs disaggregated data

Aggregation

Campaign	Site
Acme	Zombo.com
Acme	Zombo.com
Acme	Zombo.com
Acme	Nyan.cat
Brawndo	Zombo.com
Brawndo	Nyan.cat
Brawndo	Nyan.cat

Aggregation

Campaign	Site
Acme	Zombo.com
Acme	Zombo.com
Acme	Zombo.com
Acme	Nyan.cat
Brawndo	Zombo.com
Brawndo	Nyan.cat
Brawndo	Nyan.cat

Aggregation

Campaign	Site
Acme	Zombo.com
Acme	Zombo.com
Acme	Zombo.com
Acme	Nyan.cat
Brawndo	Zombo.com
Brawndo	Nyan.cat
Brawndo	Nyan.cat

Aggregation

Count	Campaign	Site
1	Acme	Zombo.com
1	Acme	Zombo.com
1	Acme	Zombo.com
1	Acme	Nyan.cat
1	Brawndo	Zombo.com
1	Brawndo	Nyan.cat
1	Brawndo	Nyan.cat

Aggregation

Count	Campaign	Site
3	Acme	Zombo.com
1	Acme	Nyan.cat
1	Brawndo	Zombo.com
2	Brawndo	Nyan.cat

Aggregation

Count	Campaign	Site	Lots	More
3	Acme	Zombo.com	...	...
1	Acme	Nyan.cat	...	...
1	Brawndo	Zombo.com	...	…
2	Brawndo	Nyan.cat	...	...

Lots of buckets

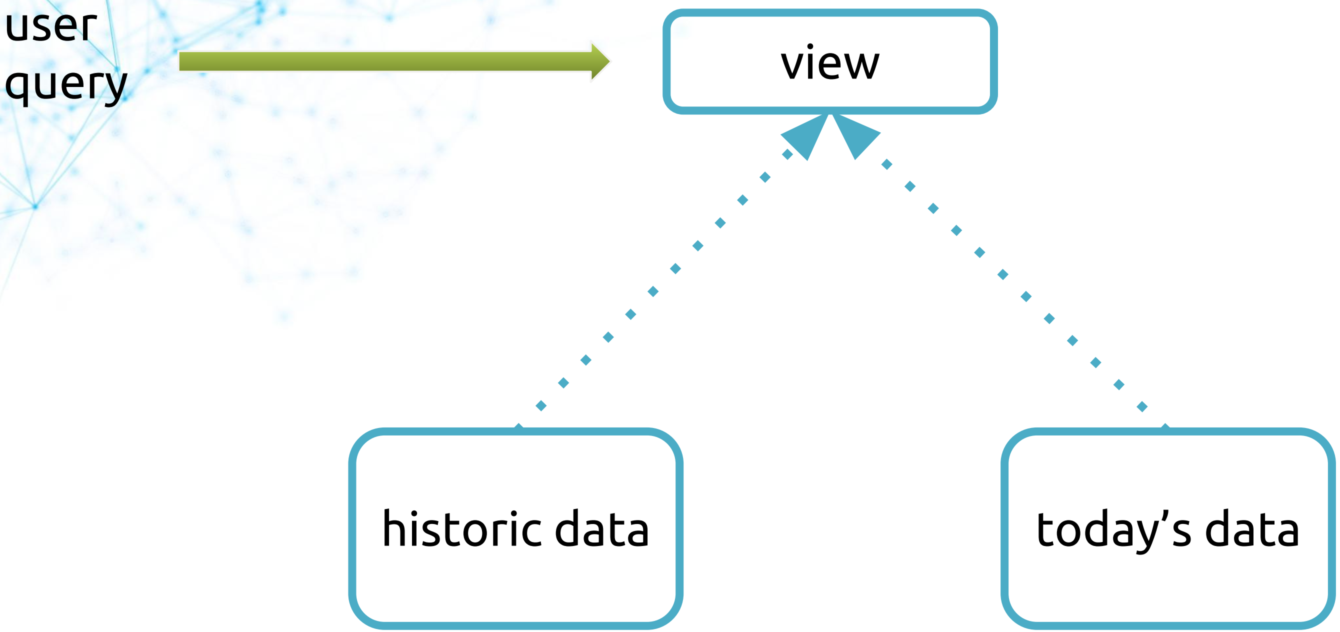
Micro-Aggregations

- Roughly 20k events per second
- Batched: window size 20s
- x7 reduction factor
- Reduces writes to db

Make America Aggregate Again

- Daily
- From ~800 million events
- Compacts to ~2 million rows
- 400x reduction
- Reduces disk usage
- Speeds up queries

Querying data



Aggregatable facts

Campaign	Site
Acme	Zombo.com
Acme	Zombo.com
Acme	Zombo.com
Acme	Nyan.cat
Brawndo	Zombo.com
Brawndo	Nyan.cat
Brawndo	Nyan.cat

Add in session ids

Campaign	Site	Session Id
Acme	Zombo.com	Wo5Meiri
Acme	Zombo.com	Xotaipu6
Acme	Zombo.com	Xu1goor7
Acme	Nyan.cat	eVai6OhS
Brawndo	Zombo.com	oiMoo7Du
Brawndo	Nyan.cat	aiSh1eej
Brawndo	Nyan.cat	rae8ieY5

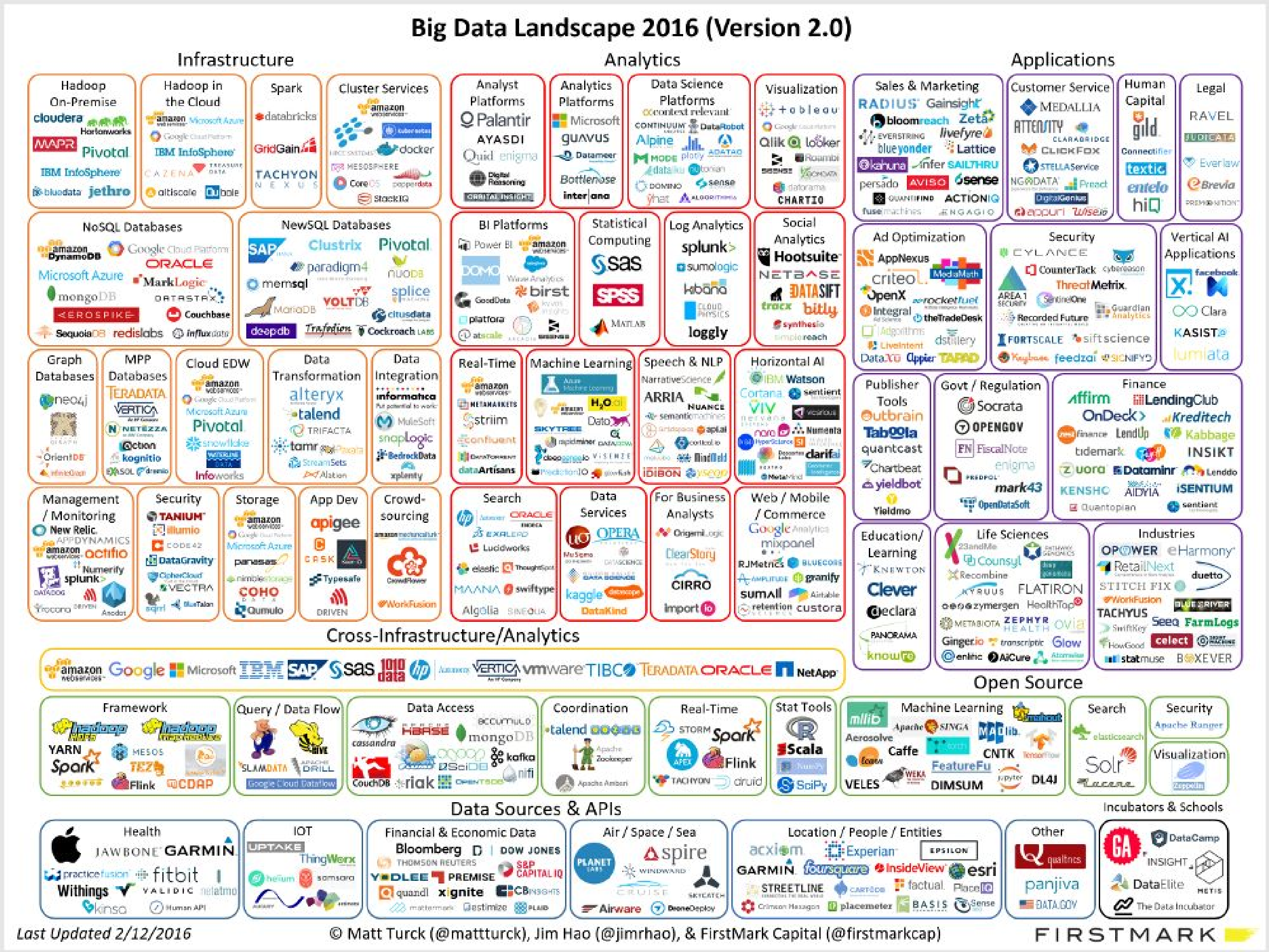
Does not aggregate well

Campaign	Site	Session Id
Acme	Zombo.com	Wo5Meiri
Acme	Zombo.com	Xotaipu6
Acme	Zombo.com	Xu1goor7
Acme	Nyan.cat	eVai6OhS
Brawndo	Zombo.com	oiMoo7Du
Brawndo	Nyan.cat	aiSh1eej
Brawndo	Nyan.cat	rae8ieY5

What next?

What next? Spikes!

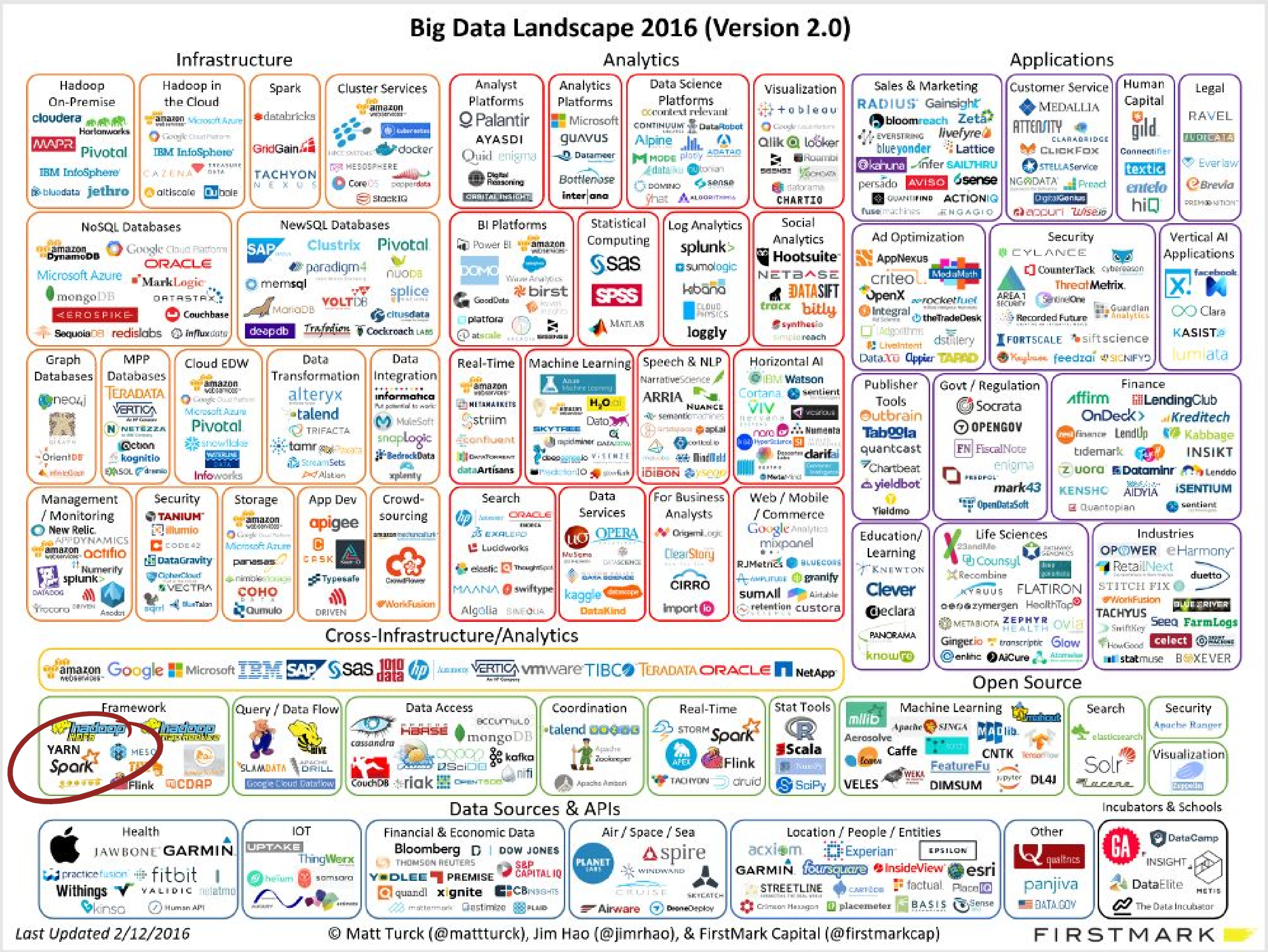
Big Data!


Big Data Landscape 2016 (Version 2.0)
Infrastructure
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
NoSQL Databases
NewSQL Databases
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Analytics
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Applications
Sales & Marketing
Customer Service
Human Capital
Legal
Ad Optimization
Security
Vertical AI Applications
Publisher Tools
Govt / Regulation
Finance
Education/ Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Incubators & Schools
Last Updated 2/12/2016
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

Big data: big choices

- Many options
- Available documentation was:
  - Academic
  - Evangelical
  - Naive/Trivial

Spark!


Big Data Landscape 2016 (Version 2.0)
Infrastructure
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
NoSQL Databases
NewSQL Databases
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Analytics
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Applications
Sales & Marketing
Customer Service
Human Capital
Legal
Ad Optimization
Security
Vertical AI Applications
Publisher Tools
Govt / Regulation
Finance
Education/ Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
YARN
Spark
Flink
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Incubators & Schools
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Last Updated 2/12/2016
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

Big data: big costs

- Infrastructure
- Language (Scala)
- Incompatible with current approach
- Performance tradeoffs

Why we could step away

- Understood our data better
- Underestimated costs
- We know our code
- We can change our code

Feedback

- Regular retrospectives
- Shared understanding of “research”
- Shared understanding of *value*

Courage

- Not afraid to try new things
- Not afraid to change direction
- Not lured by what we “ought” to do

The Shape of our Data

The Shape of our Data

Disaggregated

The Shape of our Data

Disaggregated

Unsampled

The Shape of our Data

Disaggregated

Unsampled

Real Time

Programmatic Pacing

Disaggregated

Unsampled

Real Time

Operational Debugging

Disaggregated

Unsampled

Real Time

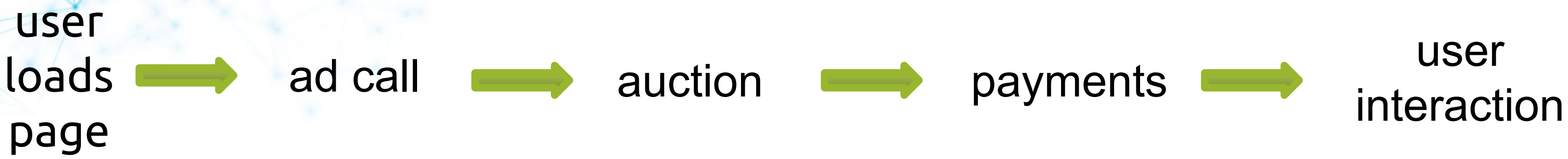
Auction Data

Disaggregated

Unsampled

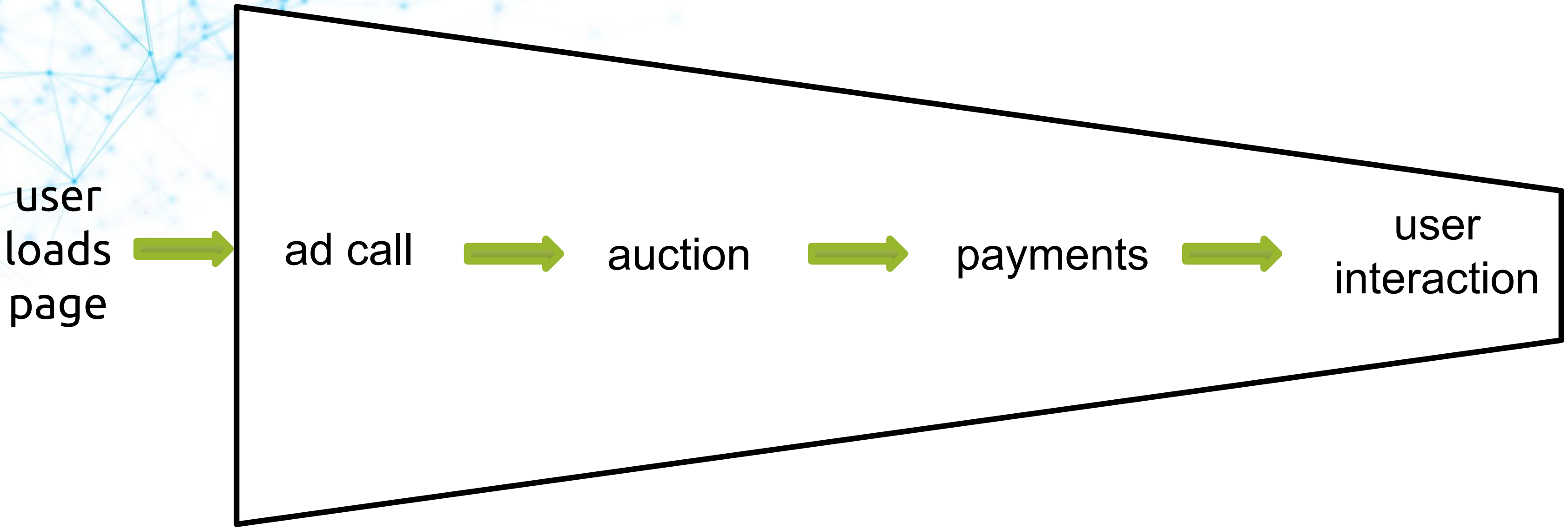
Real Time

Advertising 101



user loads page → ad call → auction → payments → user interaction

Funnel of data



user loads page

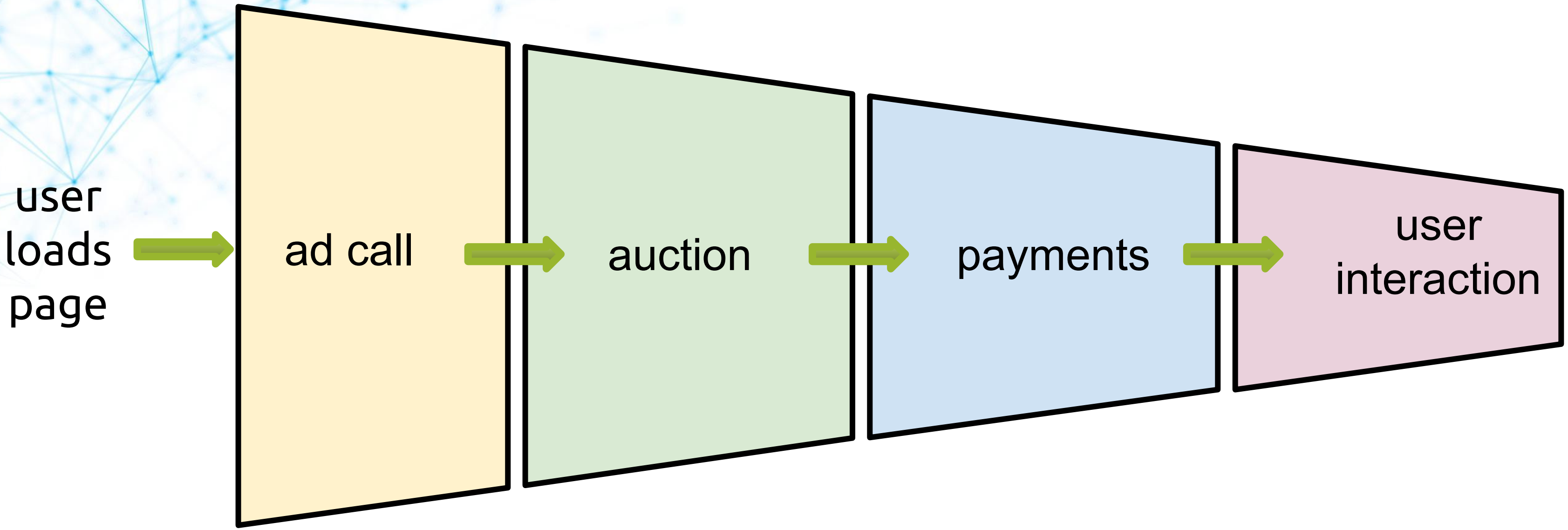
ad call

auction

payments

user interaction

Pipelines to match data shape



user loads page

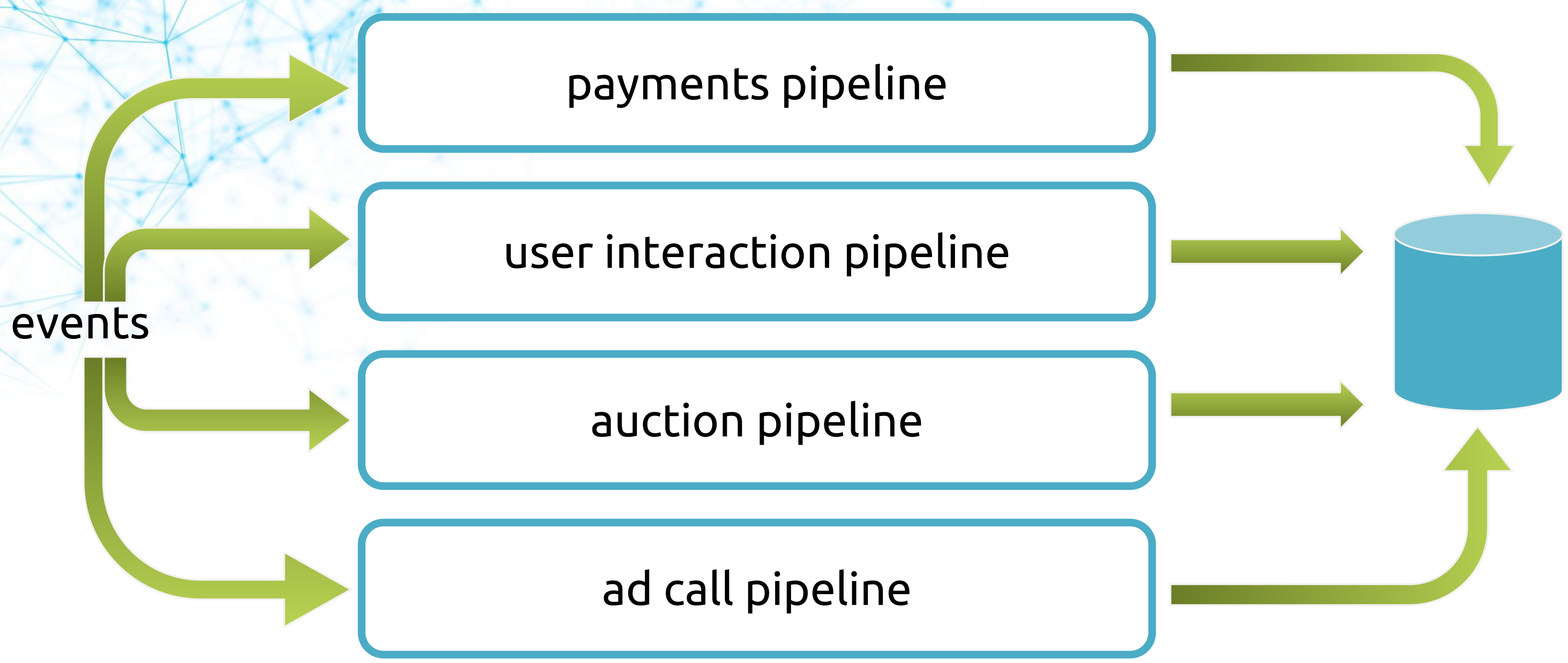
ad call

auction

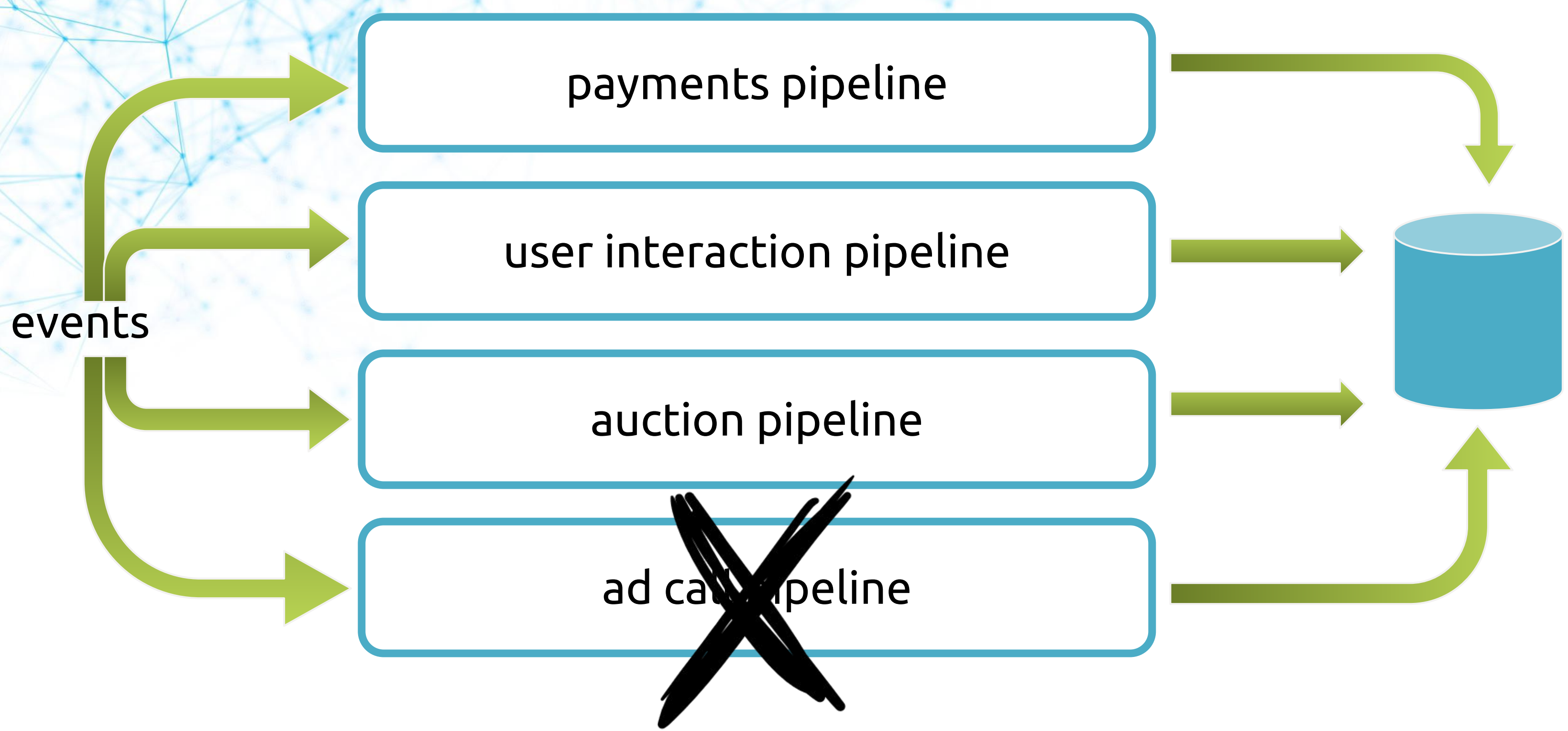
payments

user interaction

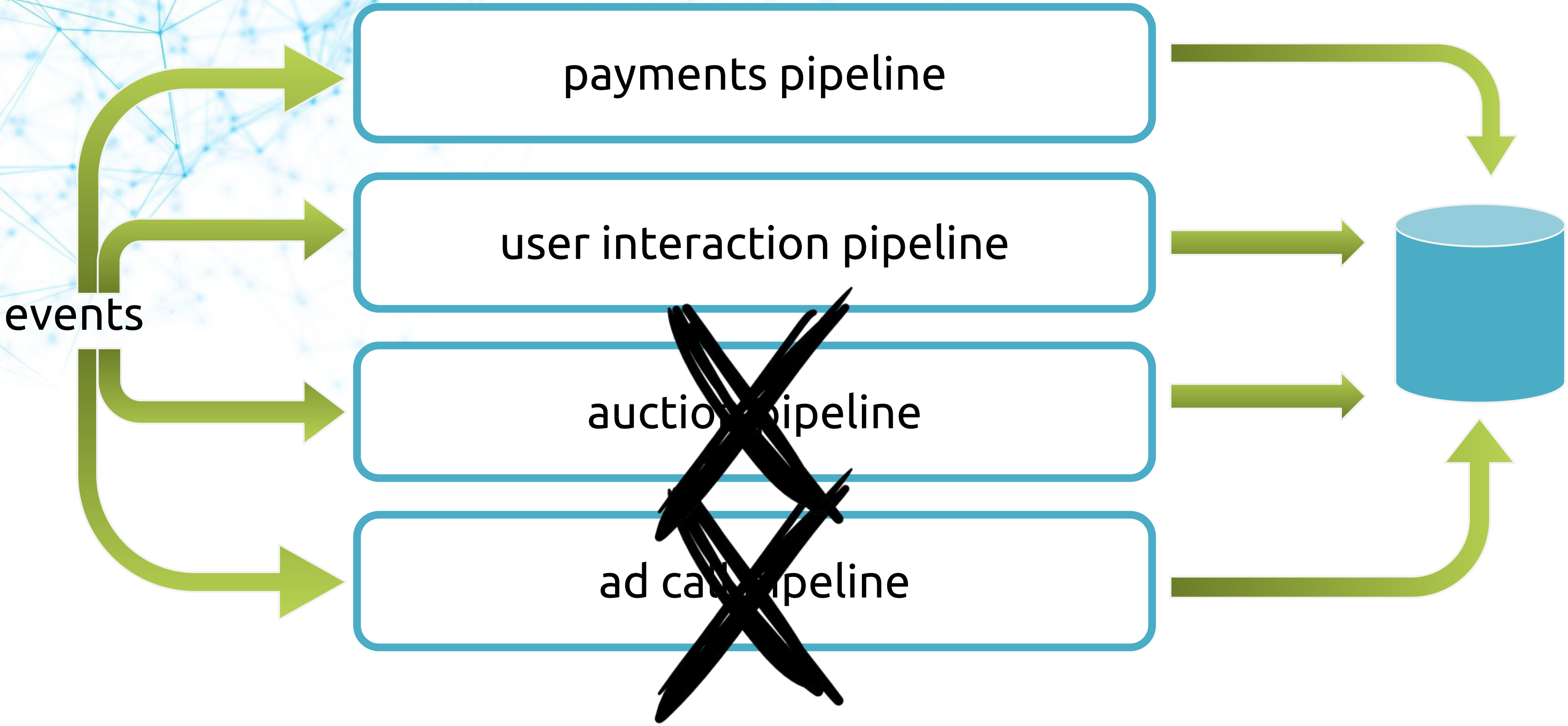
Our Actual Reporting Pipelines



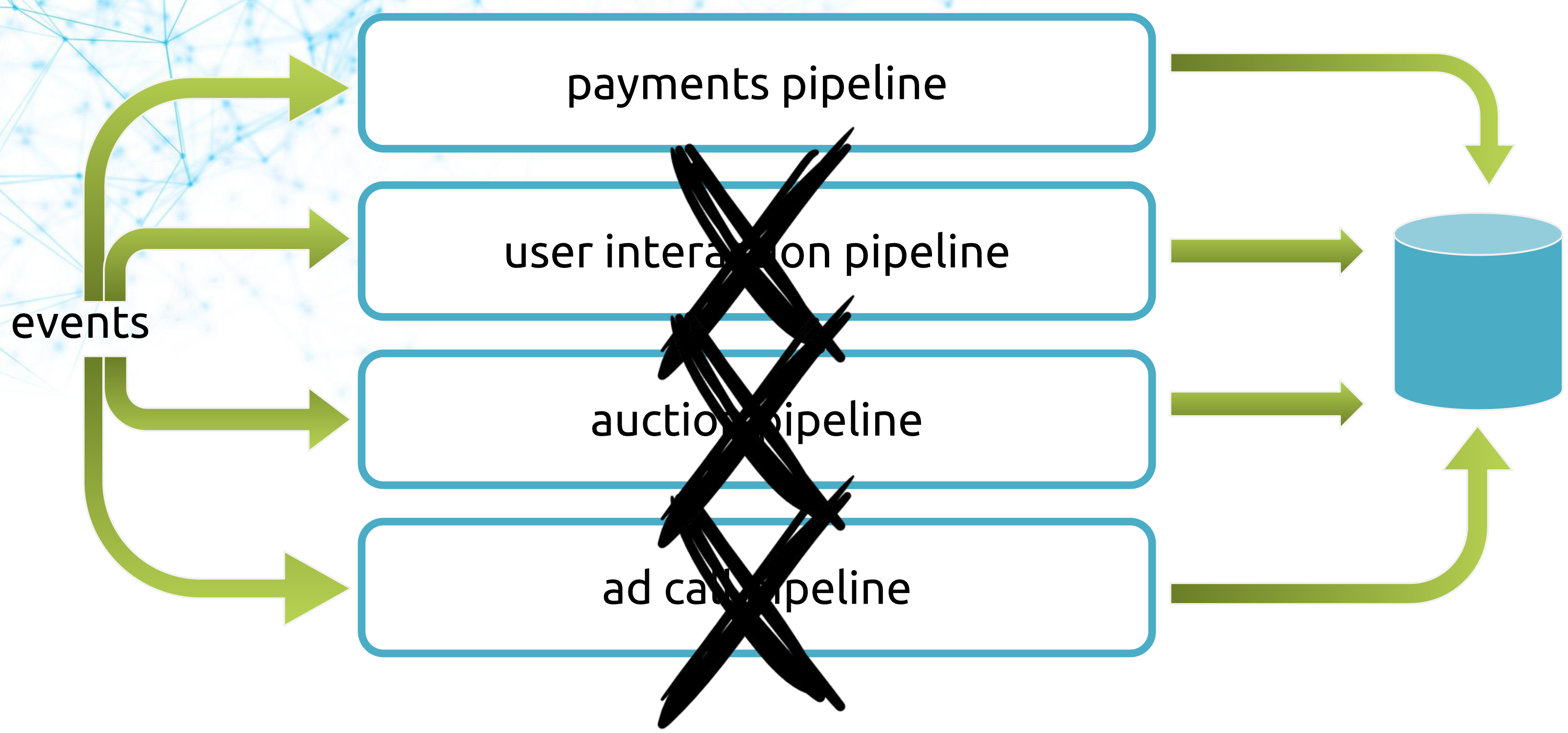
When We Get Overloaded...



When We Get Overloaded...



When We Get Overloaded...



events

payments pipeline

~~user interaction pipeline~~

~~auction pipeline~~

~~ad call pipeline~~

Ensuring real time performance

Ensuring real time performance

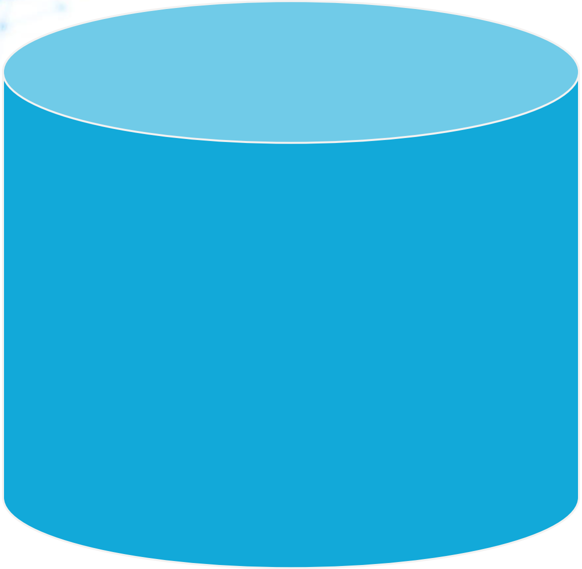
Communication

- How data was used
- Performance requirements
  - What was needed
  - What wasn’t needed
  - Hard vs soft requirements

Simplicity

- Green cards
- 10 pair-days total
- Incremental
- Separable

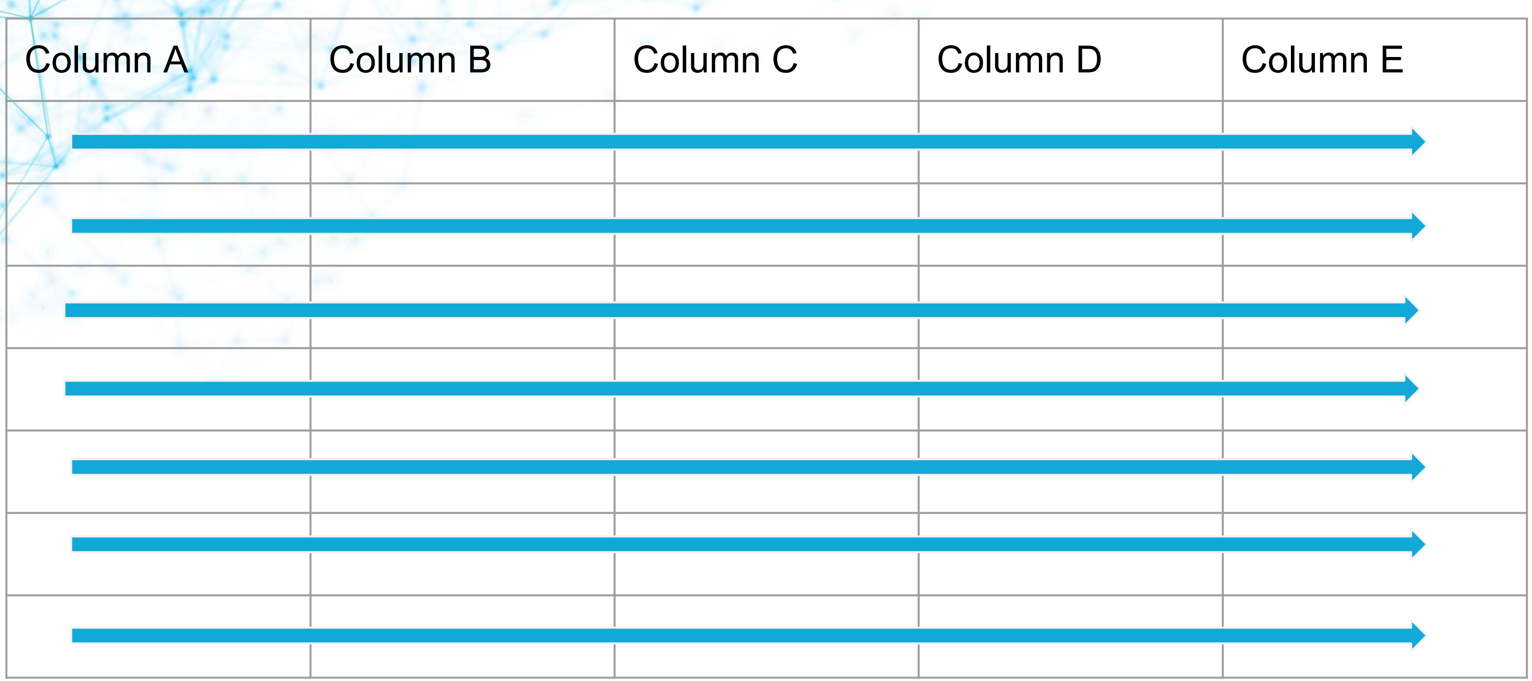
Let's talk about our databases



Row-based database

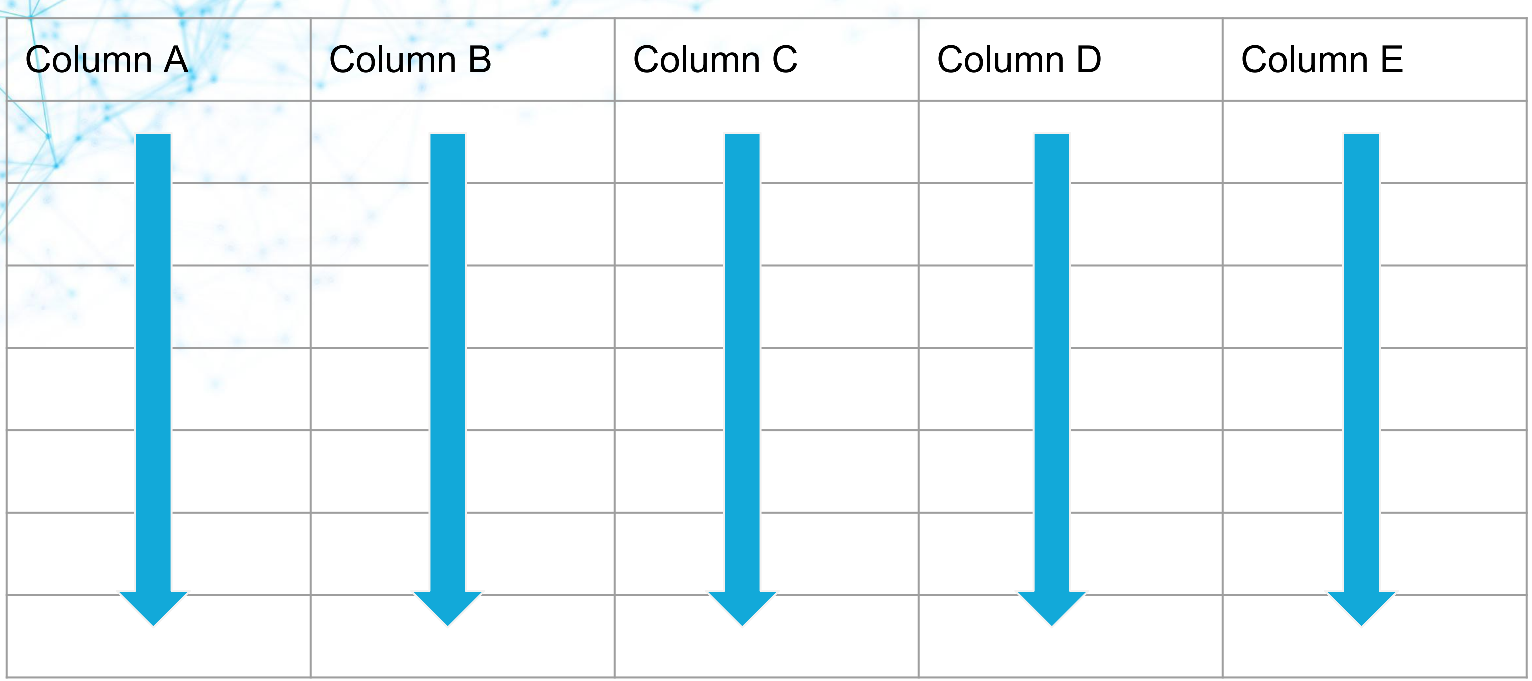
Column A	Column B	Column C	Column D	Column E

Row-based database



Column A	Column B	Column C	Column D	Column E

Columnar database



Column A	Column B	Column C	Column D	Column E

Row-based database

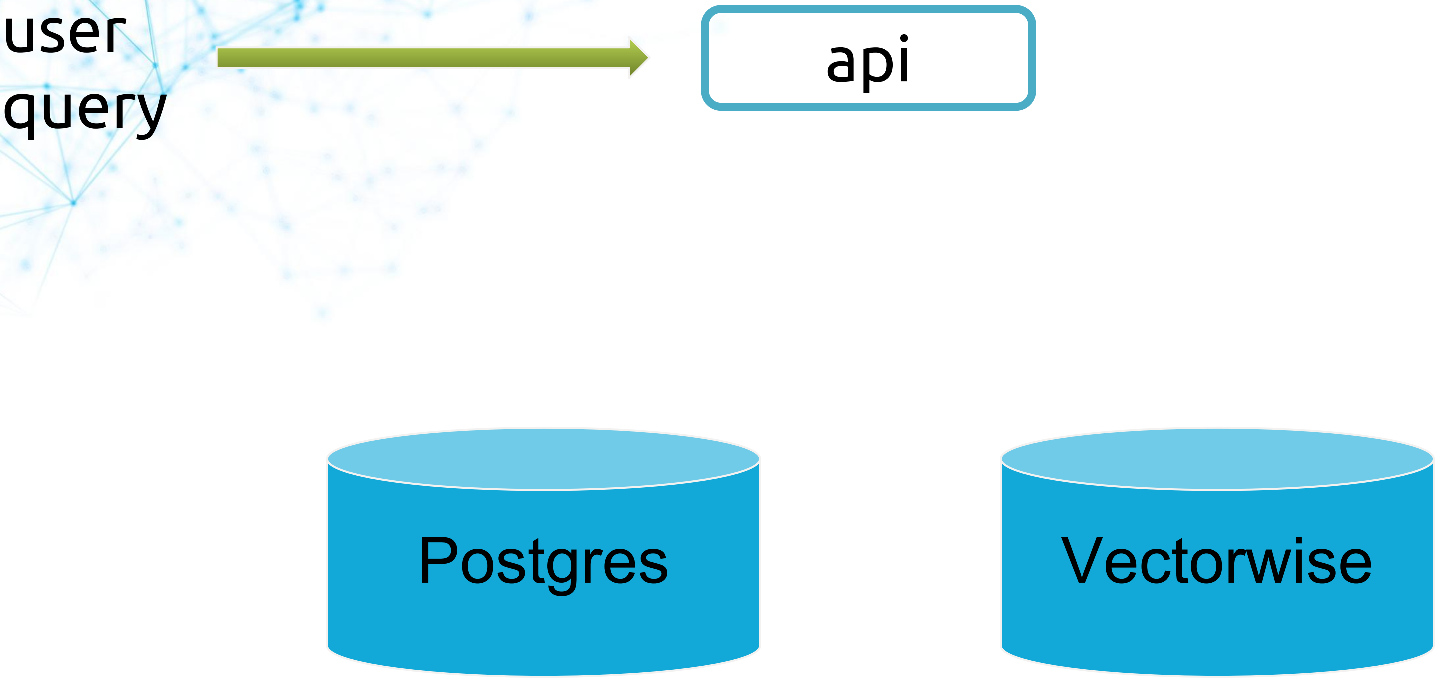
Column A	Column B	Column C	Column D	Column E

Columnar database

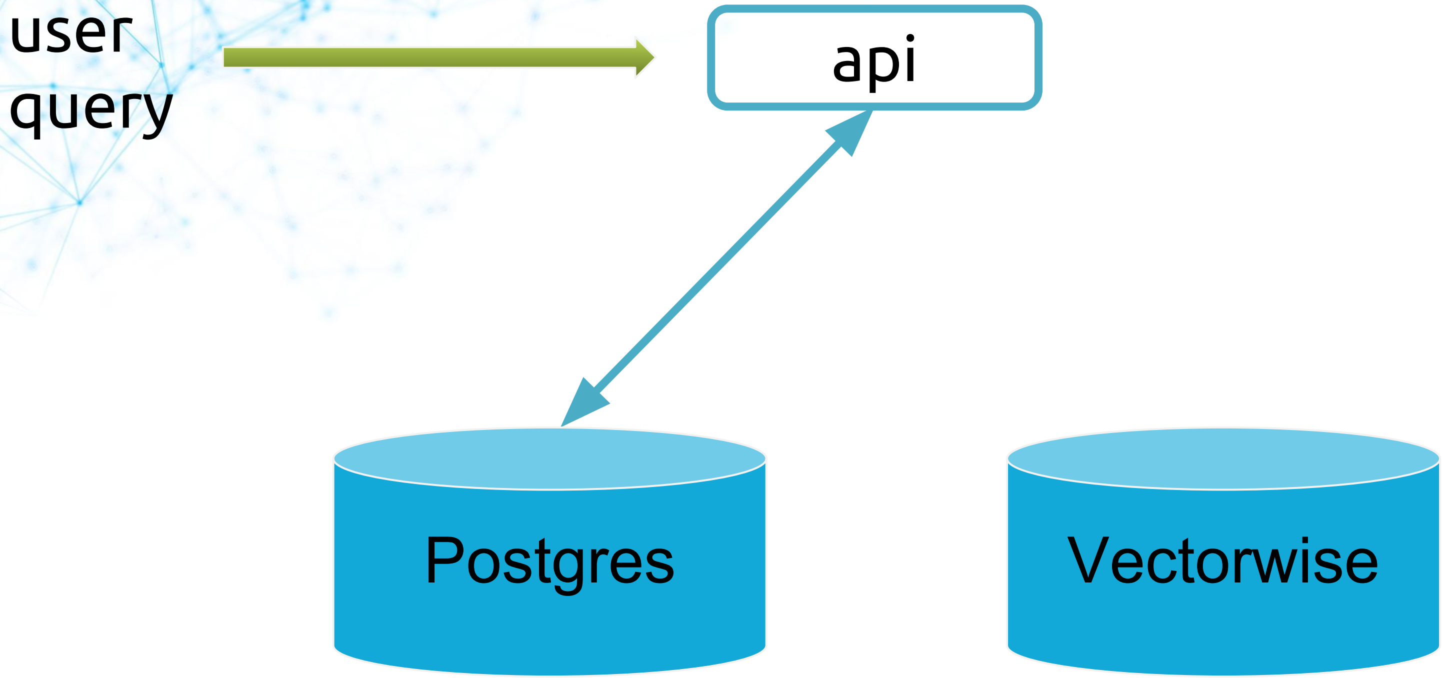
Column A	Column B	Column C	Column D	Column E
	▨			▨
	▨			▨
	▨			▨
	▨			▨
	▨			▨
	▨			▨
	▨			▨

Vectorwise or Postgres?

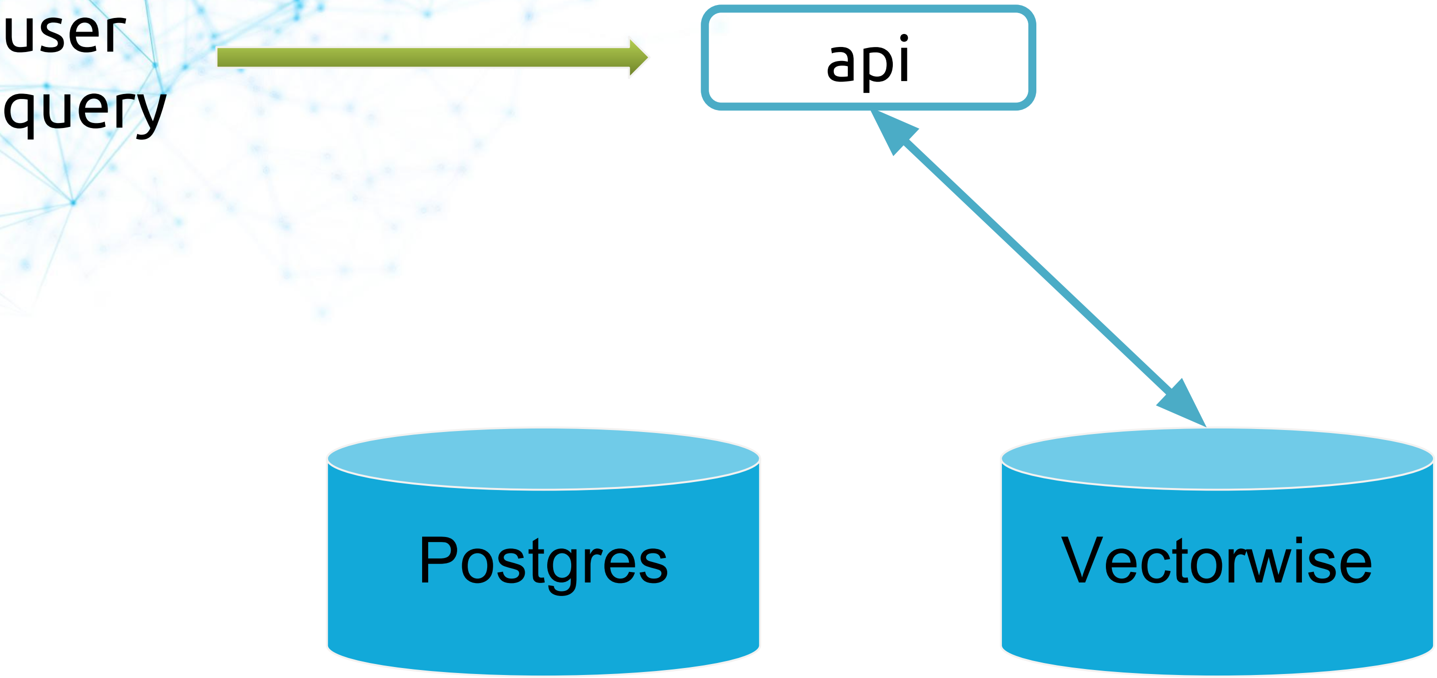
Query-based routing



Query-based routing



Query-based routing



Conclusion

Conclusion

- Simplicity
- Communication
- Feedback
- Courage

Thank you!

Questions?

(this space intentionally left blank)