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

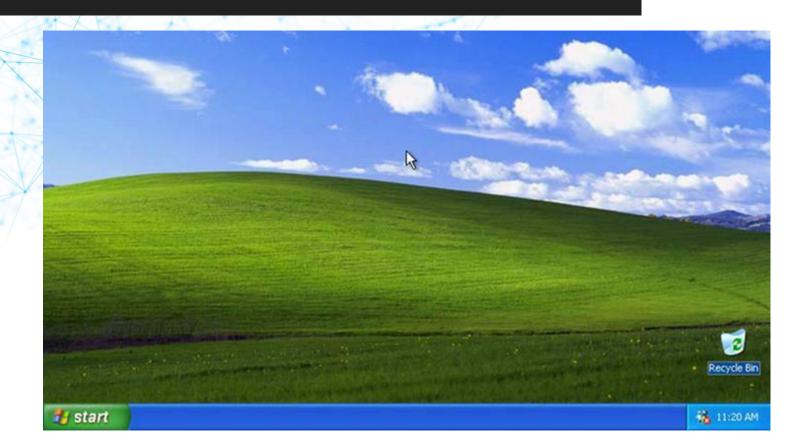


Hello. My name is...

Gel Goldsby Reporting and Data Team Lead Tom Johnson Senior Developer

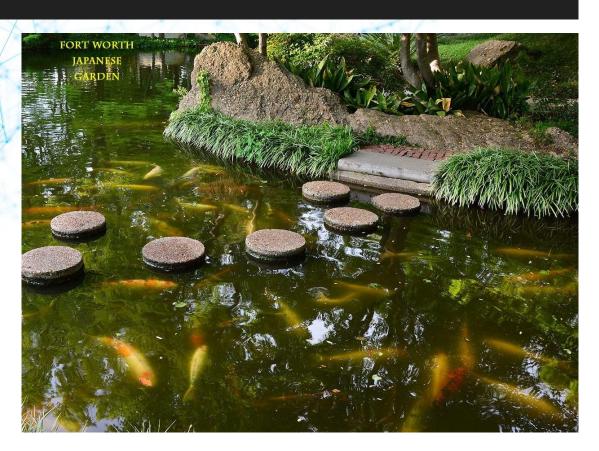


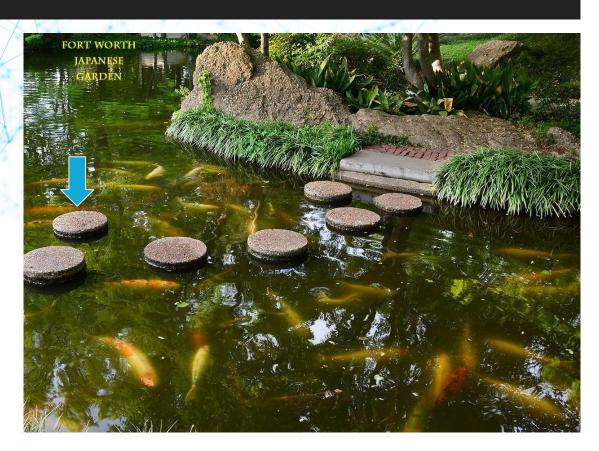
We Believe In XP

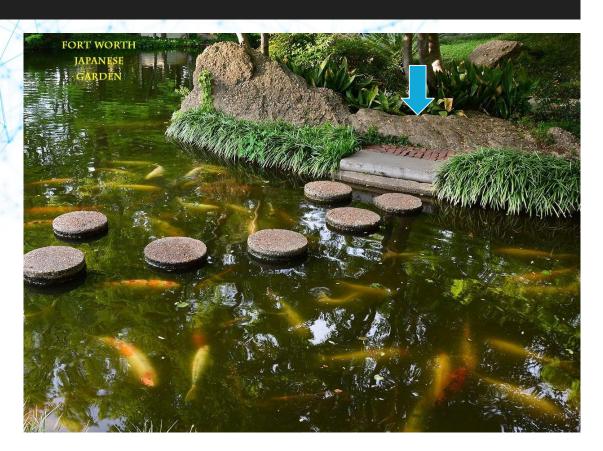


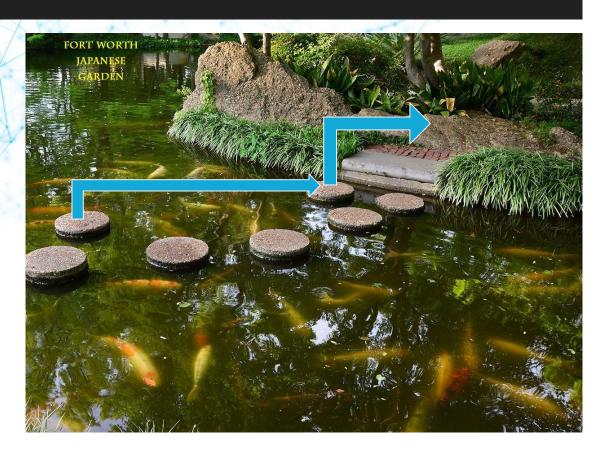
Extreme Programming Values

- Communication
- Simplicity
- Feedback
- Courage

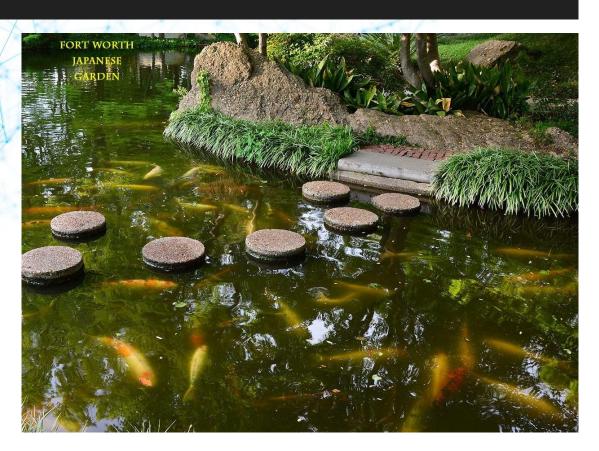




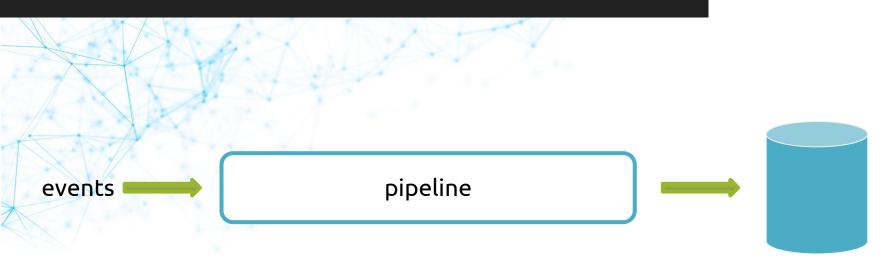




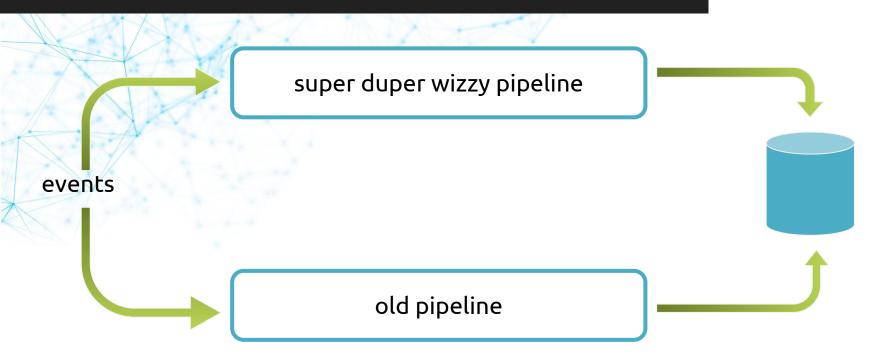




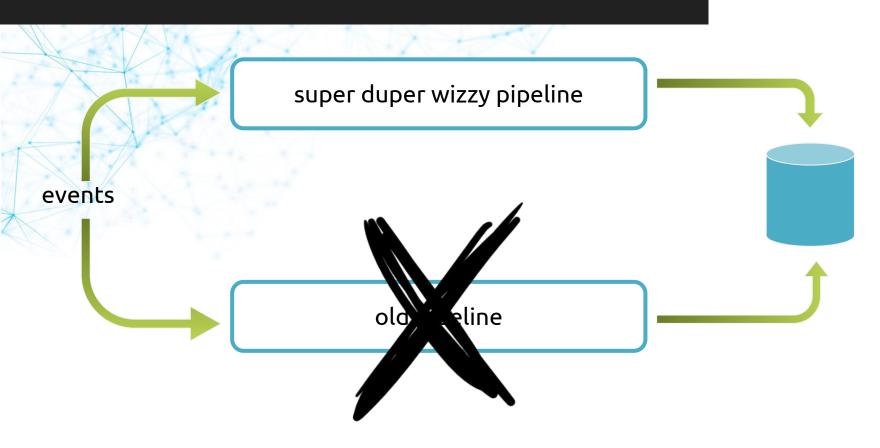
Our Reporting Pipeline



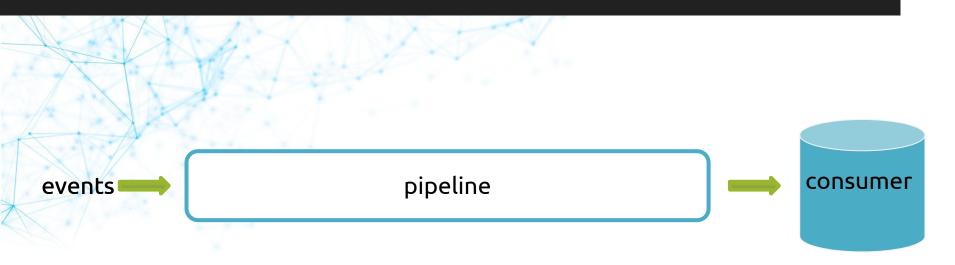
Our Reporting Pipelines



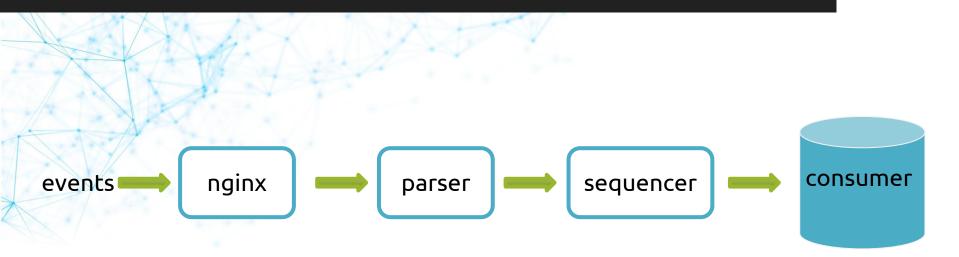
Shut It Off!



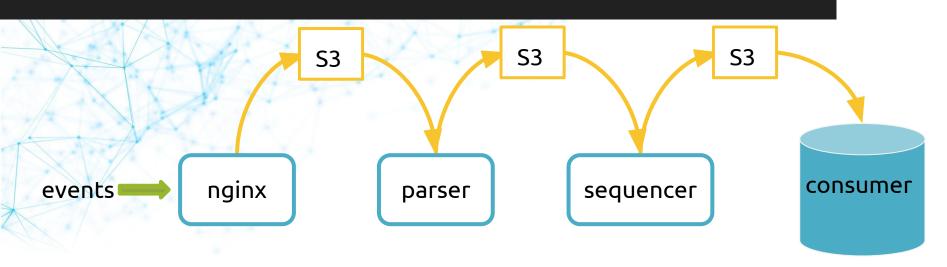
A Closer Look At Our Pipeline



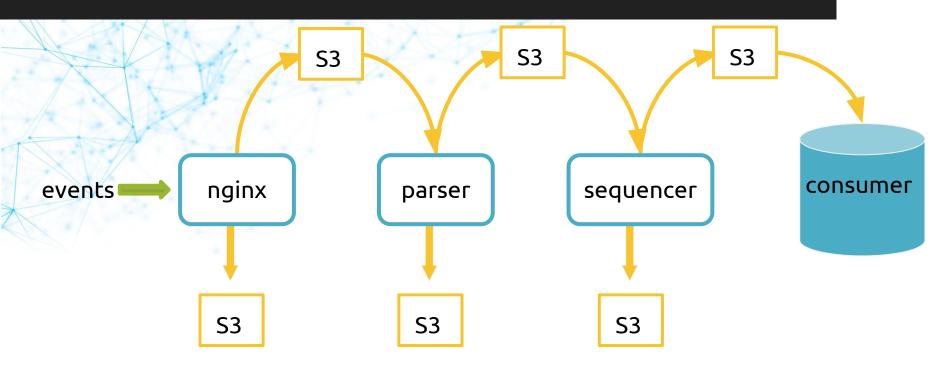
It's Not A Truck, It's A Series of Tubes

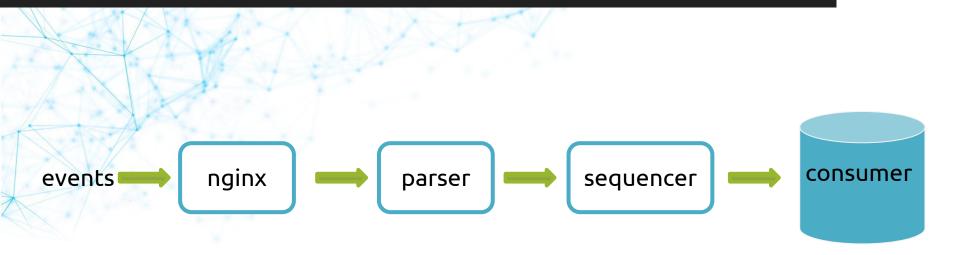


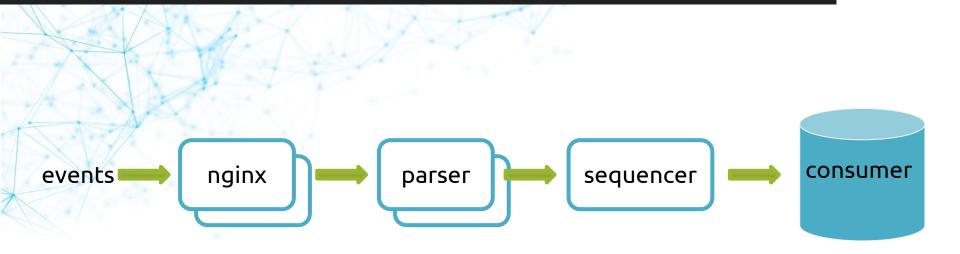
Queueing with S3

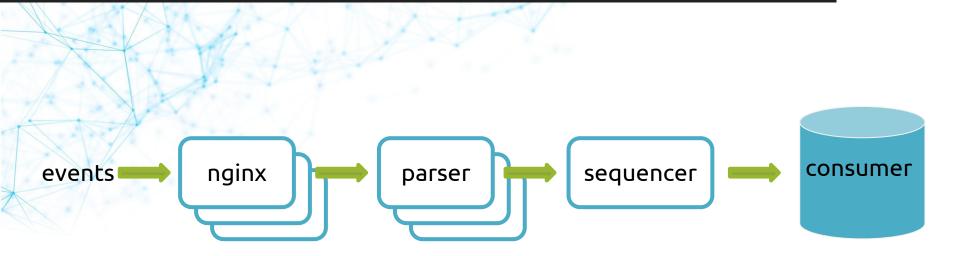


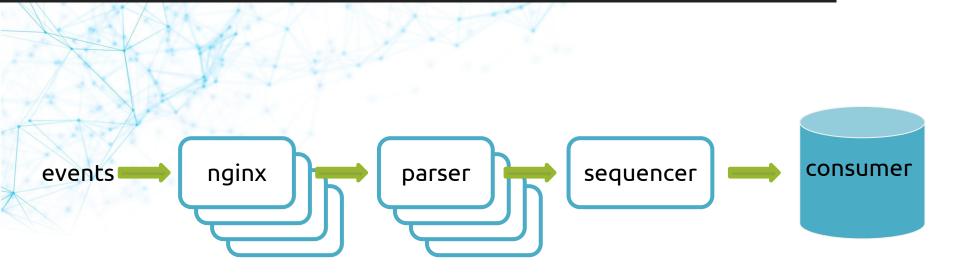
Queueing with S3



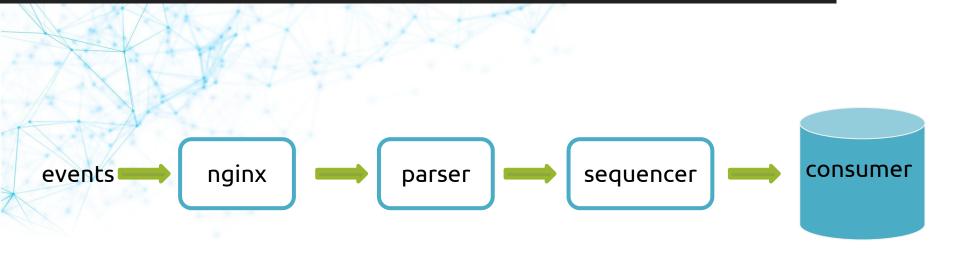




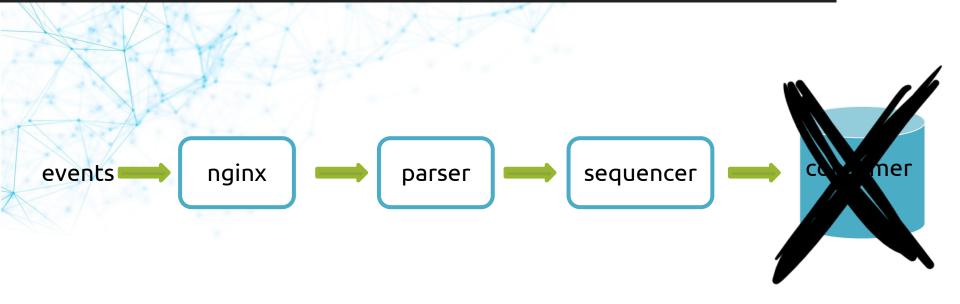




Two Writes Can Make A Wrong



Two Writes Can Make A Wrong



Christmas was saved!



 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

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

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

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

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	
	1 1 1 1 1 1 1	1 Acme 1 Acme 1 Acme 1 Acme 1 Brawndo 1 Brawndo	

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

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

Lots of buckets



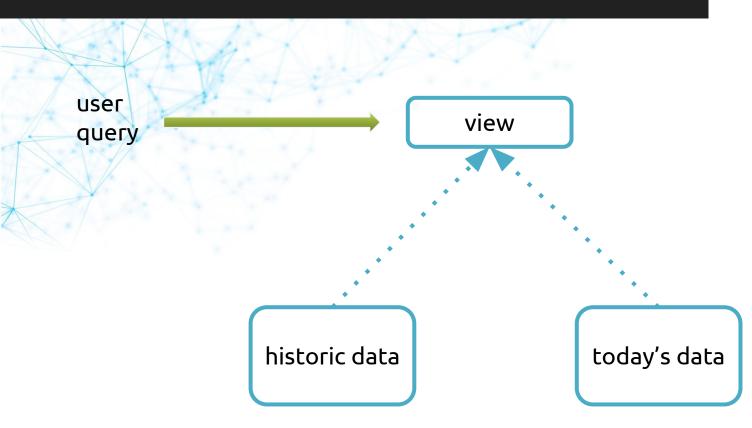
Micro-Aggregations

- 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

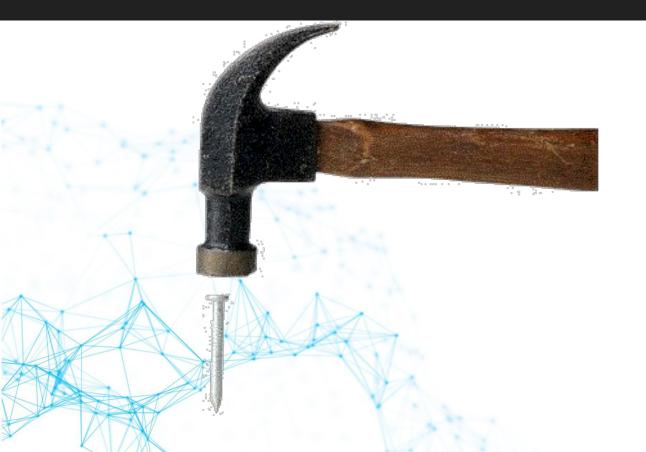
Does not aggregate well

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?

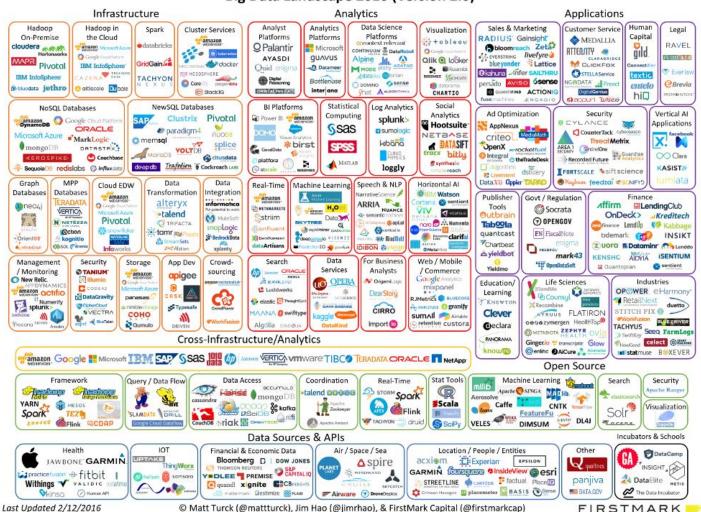


What next? Spikes!



Big Data!

Big Data Landscape 2016 (Version 2.0)



Big data: big choices

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

Infrastructure Spark! Analyst Hadoop Hadoop in Spark Cluster Services On-Premise the Cloud Platforms amazon databricks Q Palantir amazen Microsoft A Hartanwarks Google Continue AYASDI MAPR Pivotal GridGain 4 docker! IBM InfoSphere Quid enigm IBM InfoSphere TACHYON Core 05 popperdata Digital Reasoning & bluodata jethro altiscale Dibale CRRITAL INSIGHT Stack IQ NoSOL Databases NewSQL Databases **BI Platforms** Pivotal amazon Google Cloud Parton ORACLE Microsoft Azure MarkLogic memsal splice DATASTAX ■ mongo DE GoodDate **VOLTOB** Couchbase **Citusdata** platfora SequoiaDB redislabs @ Influxdate deepdb Trafedien & Cockroach Land (D atscale Graph Data Data Cloud EDW Real-Time Databases Databases Transformation Integration amazon amazon alteryx @neo4i METAMARKETS At potential to work VERTICA. *talend striim MuleSoft Pivotal N NETEZZA TRIFACTA snapLogic confluer Oction tamr *** Bedrock Data OrientDE EMPATORISENT kognitio Stream Sets dataArtisans ENASOL Odremio Infoworks Dog Alation xplenty Management App Dev Crowd-Search Storage / Monitoring TANIUM sourcing ORACLE amazon apigee X illumio New Relic. 35 EXALERD C CODE42 Lucidworks CASK A DataGravity Darwisas 👺 elastic 🖸 Thoughtion Numerify nimblest **5** Typesafe *VECTRA MAANA @ swiftype COHO Qumulo Cross-Infrastructure/Analytics amazon Google !! Microsoft TRIM SAP SSAS III (M) America VERTICA VMWARE TIBE TERADATA ORACLE II NetApp Query / Data Flow Framework Data Access Harise mongoDF cassandra PSciDe & kafka CouchDB *riak # O----Spark SLAMDATA DRILL Health Financial & Economic Data

IAWBONE GARMIN

heium a samsara

practice fusion in fit bit

©kinsa Last Updated 2/12/2016

Withings VALIDIC nelating

Big Data Landscape 2016 (Version 2.0) Analytics Applications Data Science Human Analytics Sales & Marketing **Customer Service** Visualization Legal Platforms Platforms RADIUS' Gainsight Capital MEDALLIA ** ++ableau scontext relevant RAVEL Microsof Sbloomreach Zeta gild CONTINUUM & DataRobot O Goode to A EVERSTRING livefyre **QUAVUS** CLARAGRIDGE Qlik @ looker blue yonder Lattice M CLICKFOX Connectif Datamee M MODE Roamb ®kahuna Infer SAILTHRU STELLA Service Ædatallu @tonian textic MENSE VACHOUS Bottlenase persado Aviso ósense COMINO Sense @Brevia entelo datorama. inter ana CHARTIO S GUNNEIFIND ACTIONIC DigitalGenius hi Puseimachines #NGAGIO appuri Wiselo Social Statistical Log Analytics Ad Optimization Security Vertical Al Computing Analytics Power Bl amazon splunk CYLANCE Hootsuite Applications AppNexus **G**Sas criteo MestaMent sumologic : CounterTack cybereason NETBASE Threat Metrix A DATASIFT kıbana SPSS **₹**birst OpenX perockettuet Sentine/One Recorded Future tracx bitly OO Clara) Integral theTradeDesk A MATLAB **synthesio** loggly KASIST dstillery **FORTSCALE** *siftscience Data Vi Oppier SKuybase feedzai #SICNIFYD Speech & NLP Horizontal Al Machine Learning Publisher Govt / Regulation Finance ARRIA **S** sentient **Ell Lending**Club H,O Tools Socrata Vestore Outbrain OnDeck) Kreditech Dato M (7) OPENGOV noro 🗗 🕰 Numenta @aplal Tab90la finance LendUp (Kabbage G cortical in quantcast FN FiscalNote INSIKT Personan clarifa 466 BindBelt Chartbeat Z UOra E Dataminr Challenddo PREDPOL DIBON (a vieldbot mark43 ISENTIUM Data For Business Web / Mobile "Ltt" OpenDataSoft ■ Quantopian sentient. Yleidmo Services Analysts / Commerce OPERA Google Analytica N OrigamiLogic Life Sciences Industries Education/ CHITHWAY CHICAGO OP@WER @Harmony ClearStory Learning DATASCIONO RJMetrics D BLUECO RetailNext ™ Recombine AMPLITUDE (granify CIRRO AYRUUS FLATIRON STITCH FIX 0 -Clever sumal Airtable oepezymergen HealthTap® BUTERIVER Import (6) TACHYUS Seeq FarmLog retention custora @eclara PANORAMA Gingerio * transcriptic Glow HowGood celect @#2276es Centric AiCure A. statmuse B@XEVER Open Source Stat Tools Coordination Real-Time Machine Learning Security Apache SINGA MAClib. Anache Ranger •talend @@@@@ STORM SOOK Aerosolve Scala Apache Zookeeper CNTK Visualization Solr VELES WEKE 🌃 TACHYON 🗀 dhuid Lacere Incubators & Schools Data Sources & APIs Air / Space / Sea Location / People / Entities DataCamp Bloomberg D | DOW JONES acxiem Experian ∆ spire hingWerx INSIGHT, Y-DLEE PREMISE CAPITALIO GARMIN CONSTITUTE SInsideView Cari * WINDWARD

STREETLINE CARTODO Facelle

Crimcon Hessagon D placemeter BASIS Sense

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

F Airware Tenne Deploy

quandl xignite CBNSSHTS 22 mattermade Glastimiza 880 PLAID

FIRSTMARK

panjiva

PATA GOV

A DataElite

The Data Incubator

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

Disaggregated

Disaggregated

Unsampled

Disaggregated

Unsampled

Programmatic Pacing

Disaggregated

Unsampled

Operational Debugging

Disaggregated

Unsampled

Auction Data

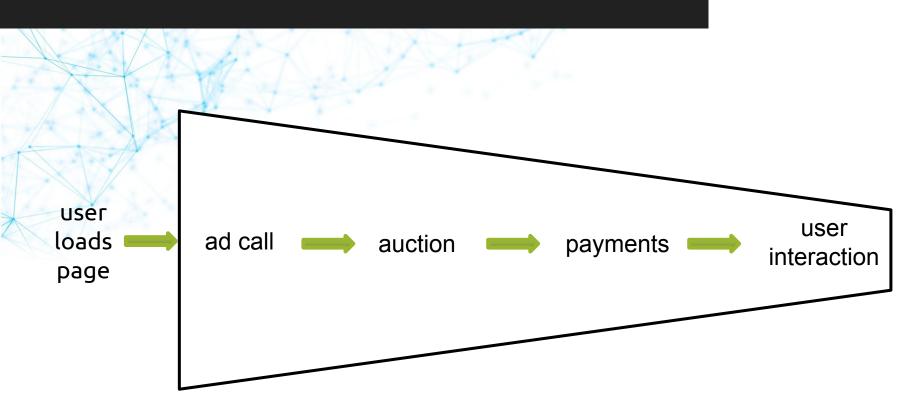
Disaggregated

Unsampled

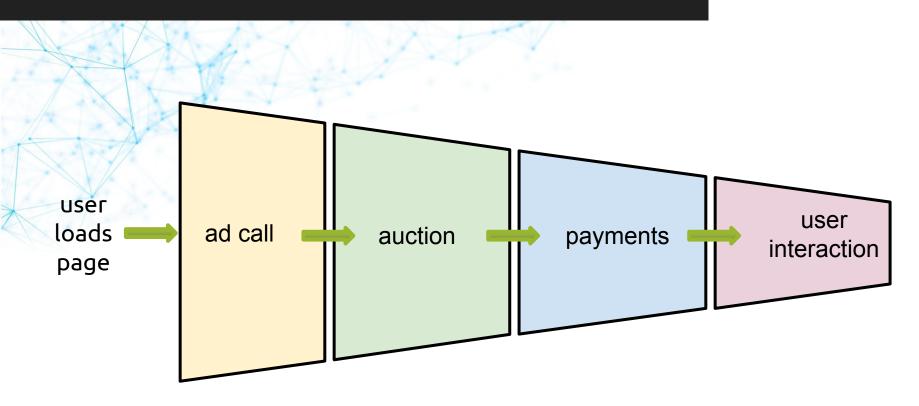
Advertising 101



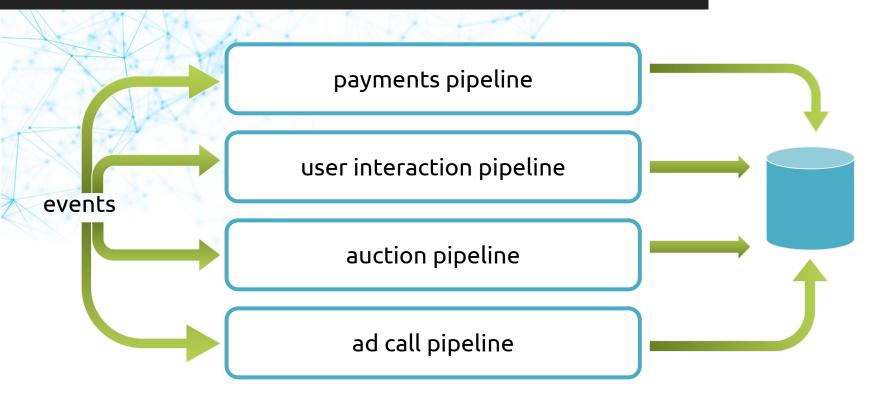
Funnel of data



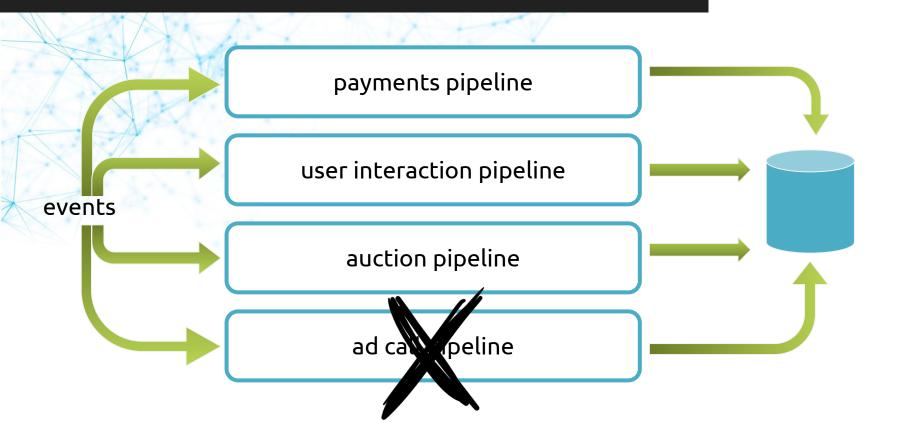
Pipelines to match data shape



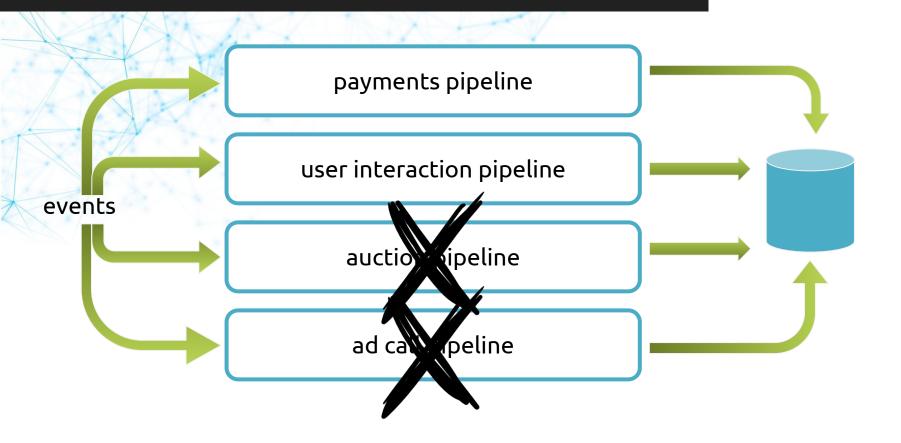
Our Actual Reporting Pipelines



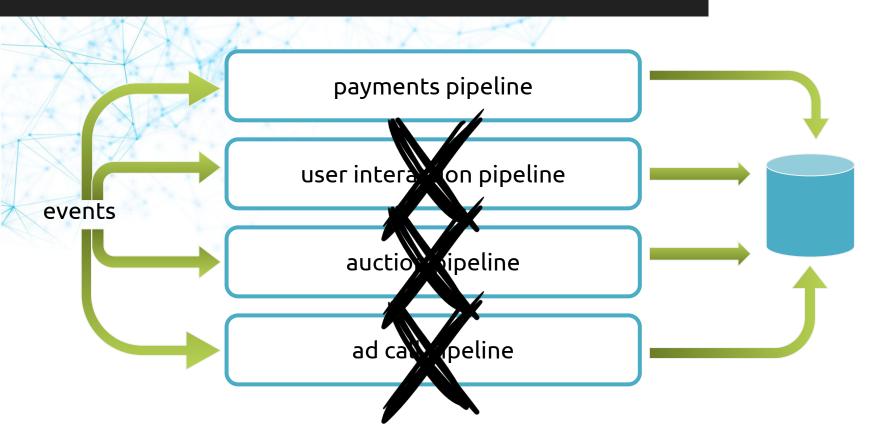
When We Get Overloaded...



When We Get Overloaded...



When We Get Overloaded...



Ensuring real time performance



Ensuring real time performance



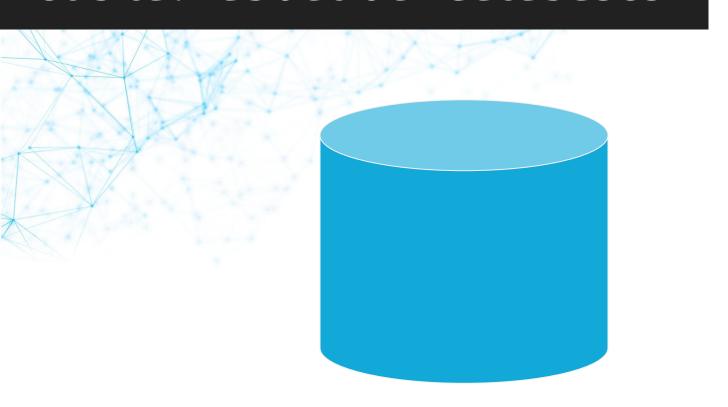
Communication

- 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



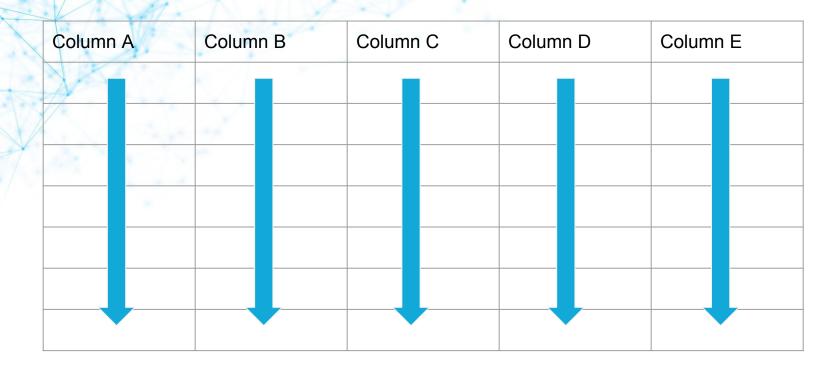
Row-based database

	Column A	Column B	Column C	Column D	Column E
1	/	.,			

Row-based database

Column A	Column B	Column C	Column D	Column E
				
-				

Columnar database



Row-based database

	Column A	Column B	Column C	Column D	Column E
1					

Columnar database

	Column A	Column B	Column C	Column D	Column E
1					

Vectorwise or Postgres?



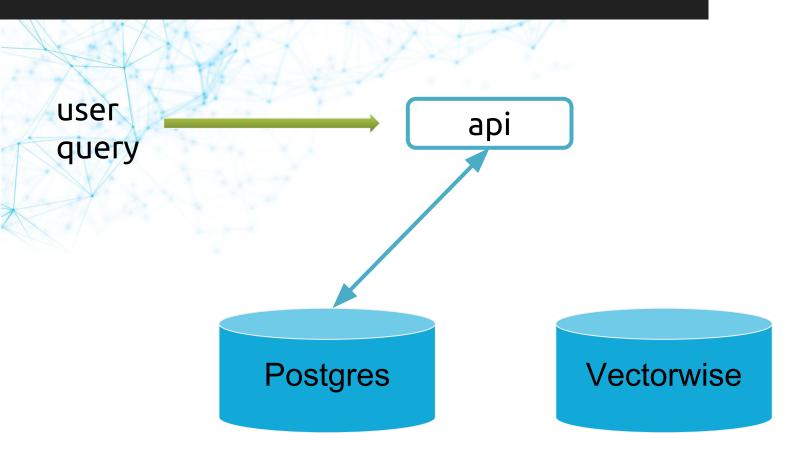
Query-based routing



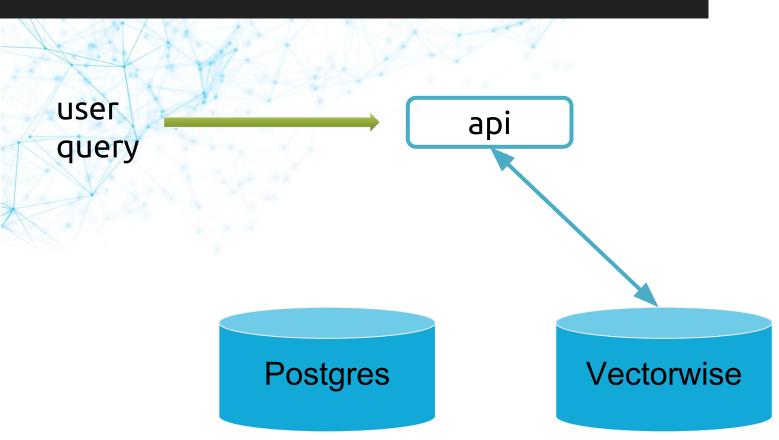
Postgres

Vectorwise

Query-based routing



Query-based routing



Conclusion

Conclusion

- Simplicity
- Communication
- Feedback
- Courage

Thank you!



Questions?

(this space intentionally left blank)