



RULES FEST



Scalability in a Real-Time Decision Platform

Kenny Shi

**Manager
Software Development**

eBay Inc.



A Typical Fraudulent Listing

Apple iPhone 4 32 GB Brand New Factory Unlocked

*Please not **BUY IT NOW** contact if you are **interested** my iPhone contact me directly by **E-mail***

- They are all 100% original Apple iPhone 4 32GB made by Apple and 100% FACTORY UNLOCKED
 - They come in original sealed boxes with all original standard accessories.
 - As you may know the iPhone 4 works with micro sim cards. I'll give you for free a Micro Sim Card cutter
 - Each phone have 1 year international Appel Care Warranty .
- I can shipp with UPS, FedEx or DHL with Air Express services (24H to 48H delivery) with tracking number.
- I have in stock more than 20 iPhone. The MINIMUM order is 2 iPhones
- I HAVE BOTH COLORS : BLACK AND WHITE

BUY NOW 2 x Apple iPhone 4 32 GB = 1000 USD BUY NOW
BUY NOW 3 x Apple iPhone 4 32 GB = 1500 USD BUY NOW
BUY NOW 5 x Apple iPhone 4 32 GB = 2000 USD BUY NOW
BUY NOW 10 x Apple iPhone 4 32 GB = 3500 USD BUY NOW

IMPORTANT !!!

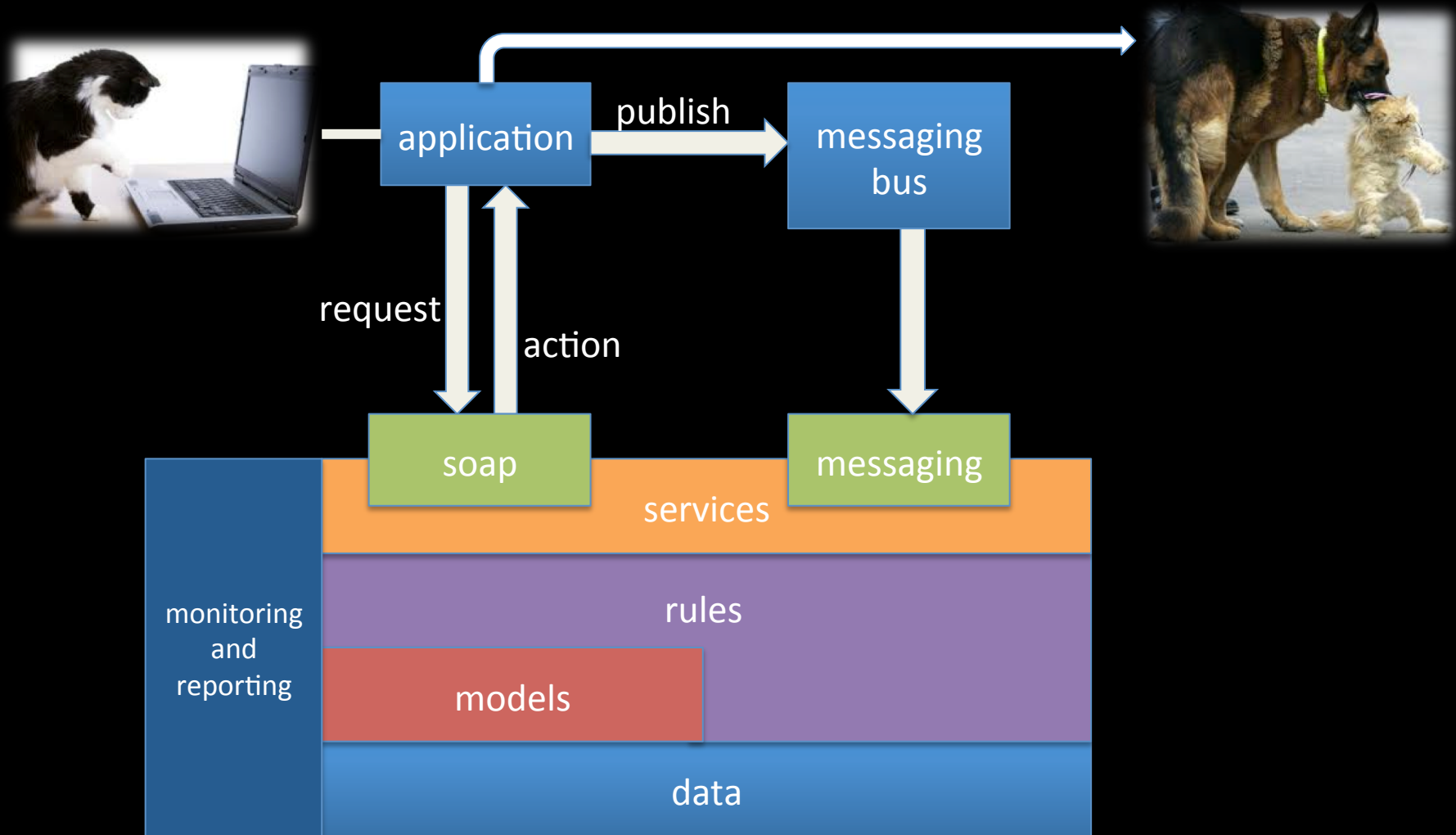
NOTE: This listing is restricted to pre-approved bidders or buyers only. Contact me before bidding or your bid will be cancelled!!!
PLEASE CONTACT ME PRIOR YOU BID USING THIS E-MAIL :

kollenrudy@gmail.com

There is No Reserve!For more info about payment and shipping please contact me.
I cannot get in touch with you if you use " Ask a seller question " option ,
so please contact me directly using this E-MAIL :

kollenrudy@gmail.com

fraud detection architecture – sync vs. async



agenda

- real-time (faster)
- scalability (more)

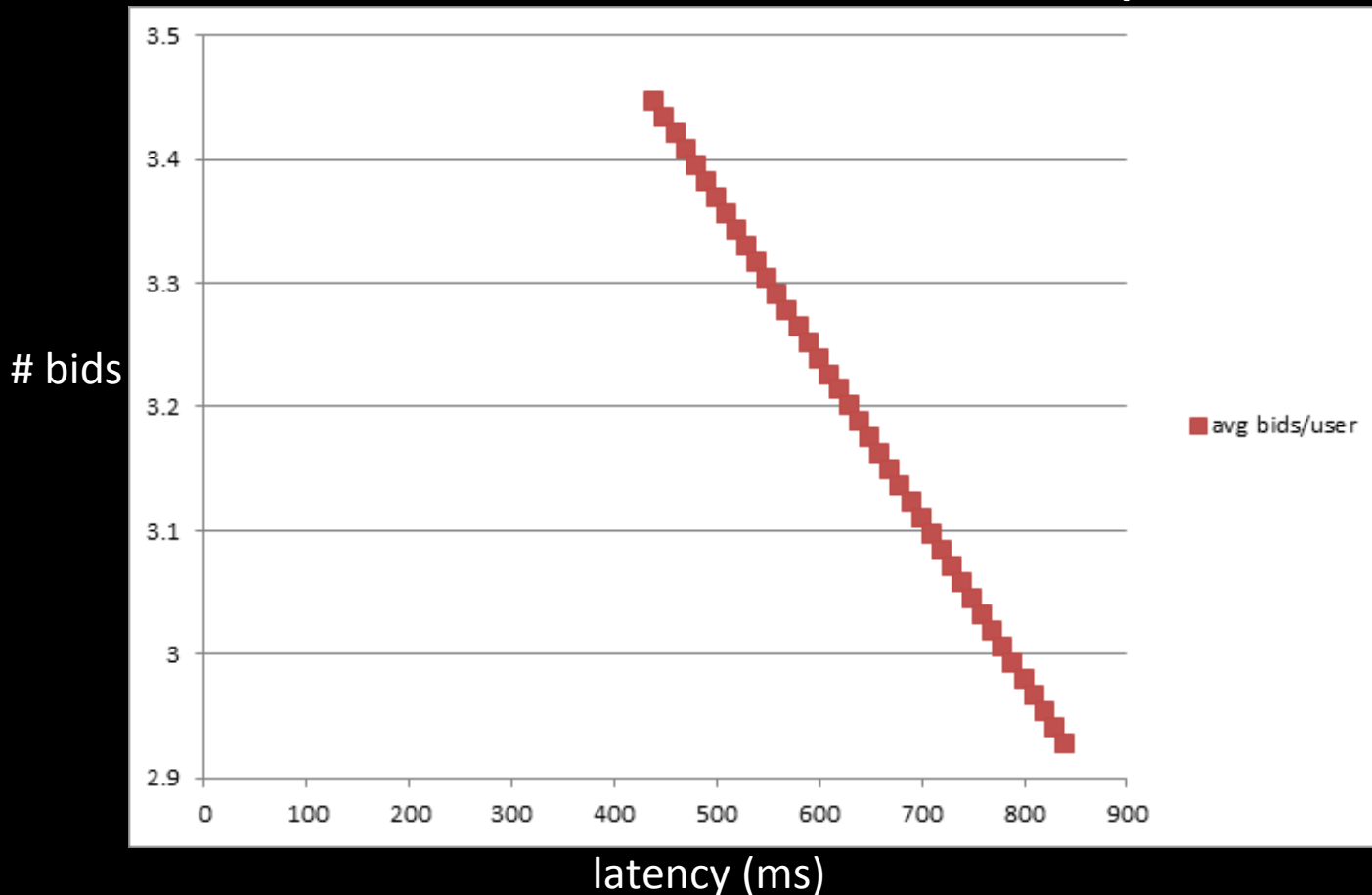


why real-time at ebay?

- mandated by business model
 - auctions
 - sign-in
- customer experience
 - response time vs. hourglass  
 - reputation of a product
- revenue impact
 - pre-transaction vs. transaction

study of revenue vs. latency

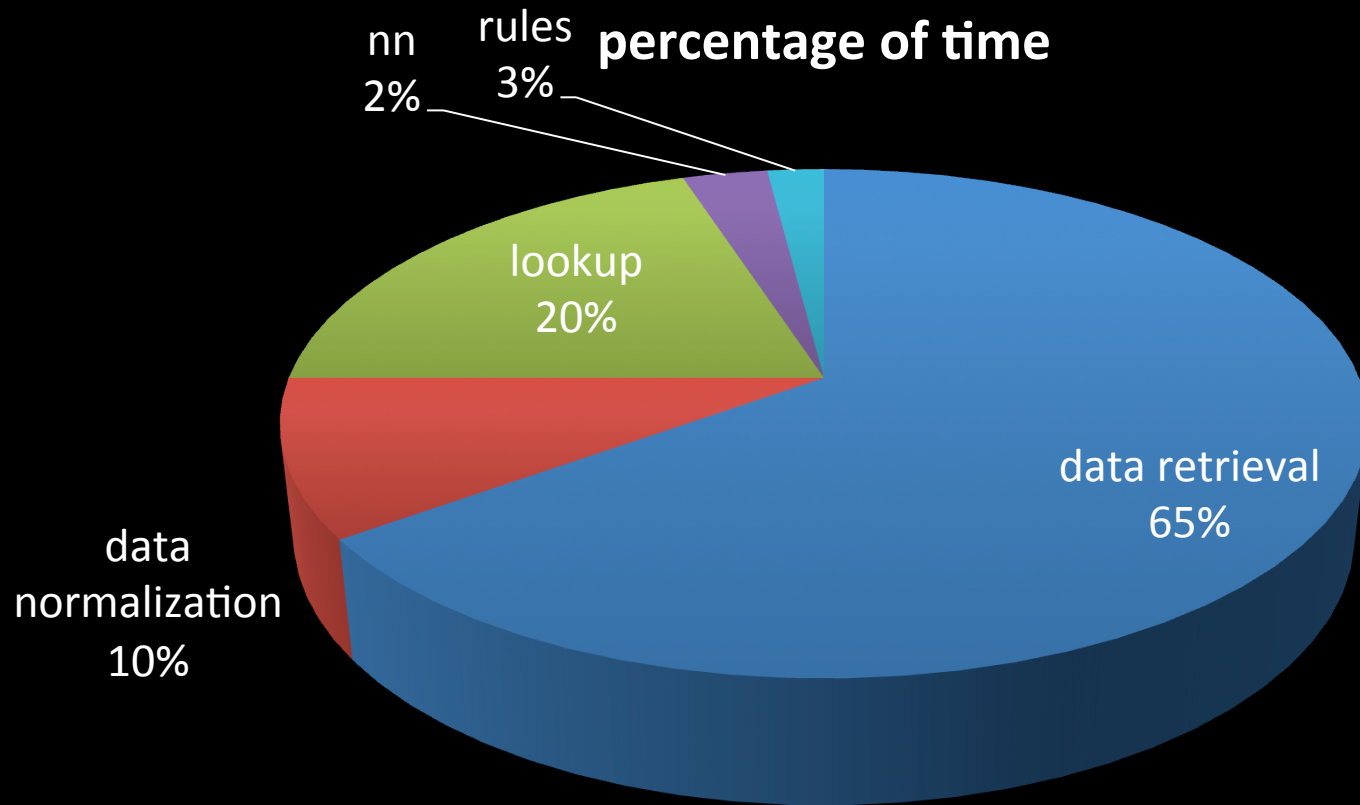
- number of bids / user vs. latency in ms



what's included in a decision?

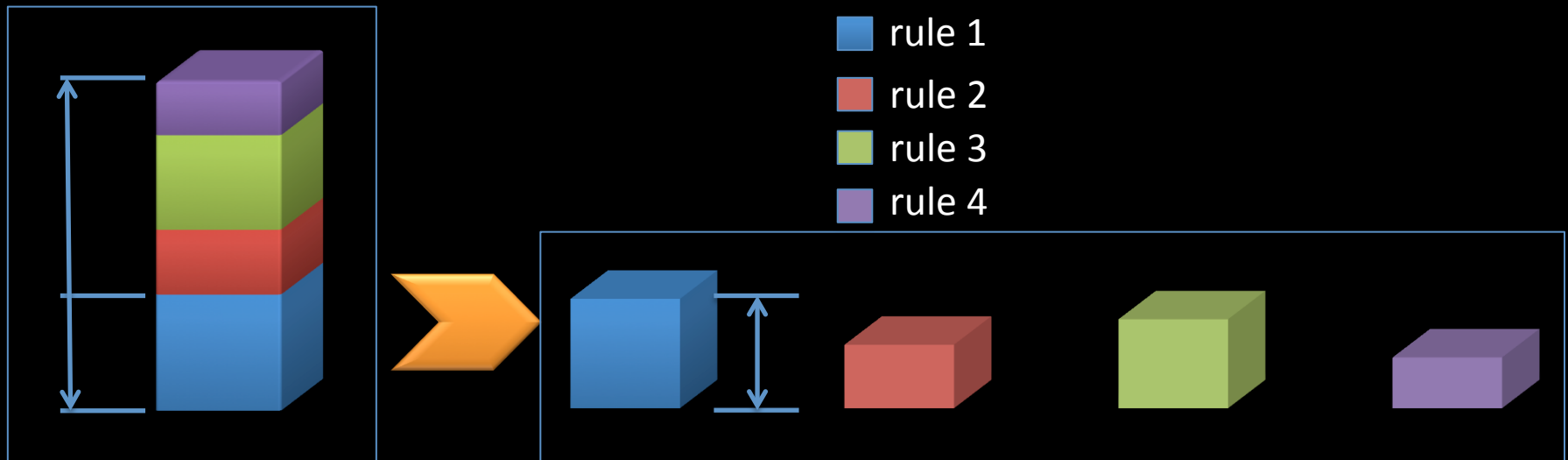
- data preparation
 - data retrieval
 - lookup data
 - data normalization / transformation
- predictive analytics models
 - neural networks
- business rules

what's taking all the time?



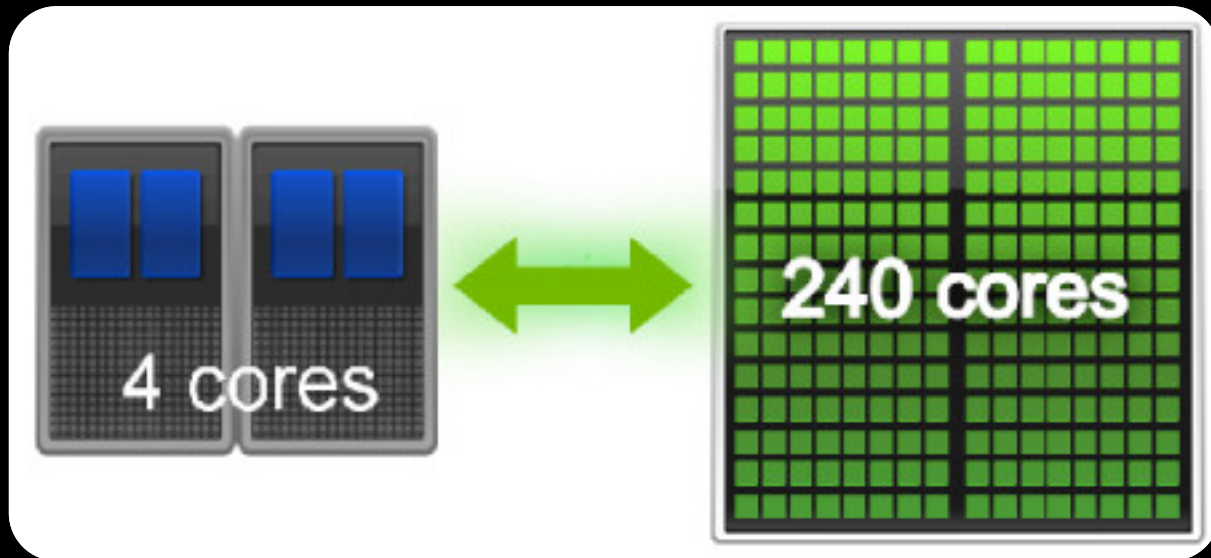
Opportunities for low latency #1

- rules are cpu-intensive and need some working memory
 - multi-core and distributed computation
- *note: we use stateless sequential evaluation



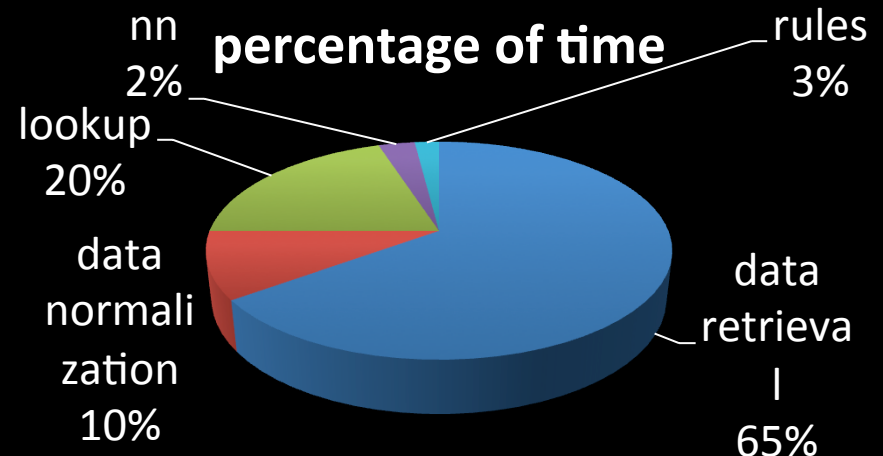
opportunities for low latency #2

- neural networks are vector and matrix computation intensive
- CUDA and GPU
 - cuda-convnet: <http://code.google.com/p/cuda-convnet/>
 - FANN: http://leenissen.dk/fann/html_latest/files2/gpu-txt.html



opportunities for low latency #3

- database calls are latency killers
 - ~6ms/call (oracle, pk hit, within datacenter, fiber-optics wired)
 - 25ms additional network latency SJC <-> PHX
- biggest bang for your buck!



data collection latency

- can we reduce database calls?
 - goal: fewer roundtrips, lower latency
- can we reduce fetch size?
 - goal: avoid large I/O transmitted
- can we parallelize database calls?
 - goal: maximize throughput of data moving
- can we eliminate database calls?
 - goal: data at your fingertips

reduce number of db calls

- can we sacrifice less important data?
 - define less important 😊
 - variables that do not trigger rules?
 - what if the variable is always true?
 - if
 - seller has credit card on file and
 - seller has no outstanding balance
 - Then ...
 - (what if prerequisite of selling is having credit card on file?)

reduce number of db calls

- excessive logic:
if
 number of transactions in 90 days > 100 and
 number of transactions in 60 days > 100
then...
- conflicting logic:
if
 user registered < 90 days ago and
 user has no activities more than 180 days
then...

reduce number of db calls

- pre-fetch
 - do we need all the prefetched data?
 - find data not used by decisions and stop loading
 - what if new rules need them?
- lazy load with in-memory caching
 - short-circuiting:
 - if a and b then... -> if a is false, b is not needed.
 - if cheap condition and expensive condition then...
 - but is cheap condition really cheaper?
 - rule 1: if expensive condition then...
 - rule 2: if cheap condition and expensive condition then...
 - challenge: difficult to estimate cost of decisions

reduce fetch size

- pre-aggregation of data
 - raw data:

```
transactions = get history (6 months);  
for each transaction in transactions  
    total value += transaction.value;
```
 - aggregated data:

```
total value = get aggregated value (6 months, today);
```
- aggregated offline, or, lazily populated
- what if staleness can't be tolerated?

reduce fetch size

- hybrid mode (aggregated data + raw data)

total value up-to-yesterday = get aggregated value (6 months ago, yesterday);

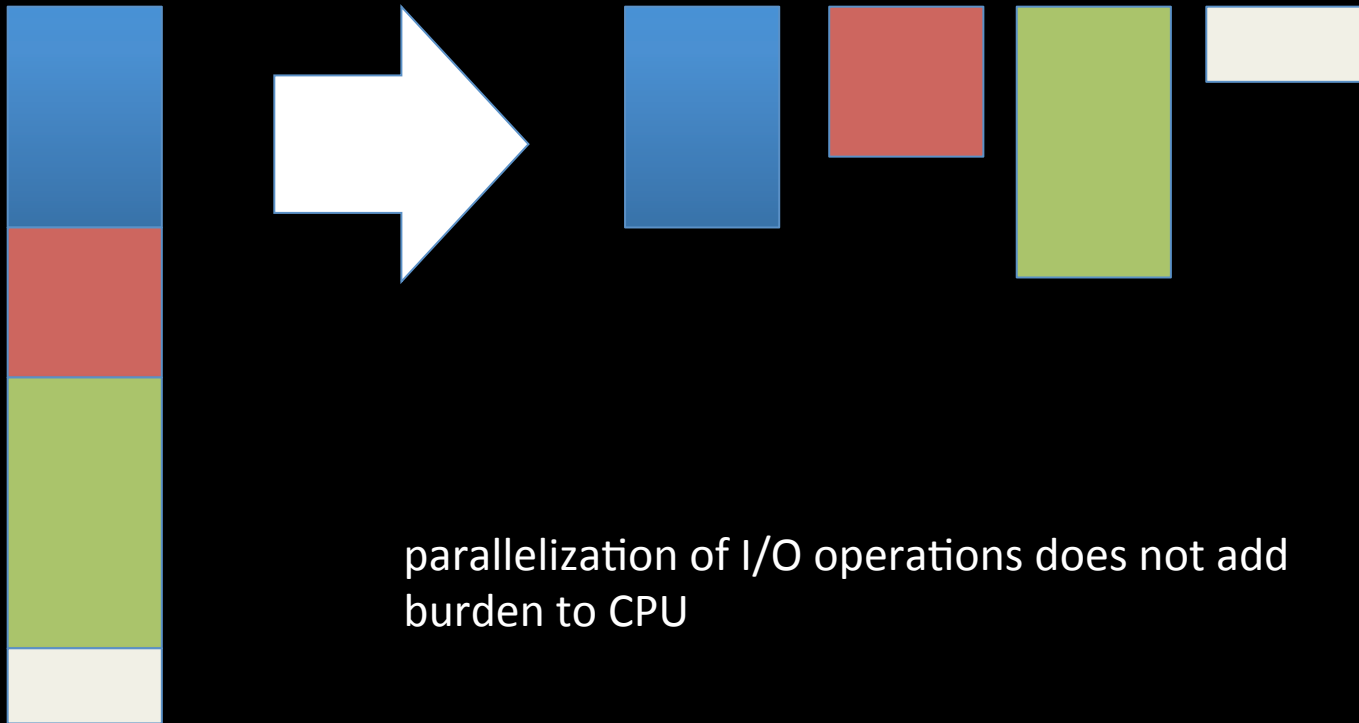
transactions today = get transactions today;

for each transaction in transactions today

total value today += transactions today.value;

total value = total value up-to-yesterday + total value today

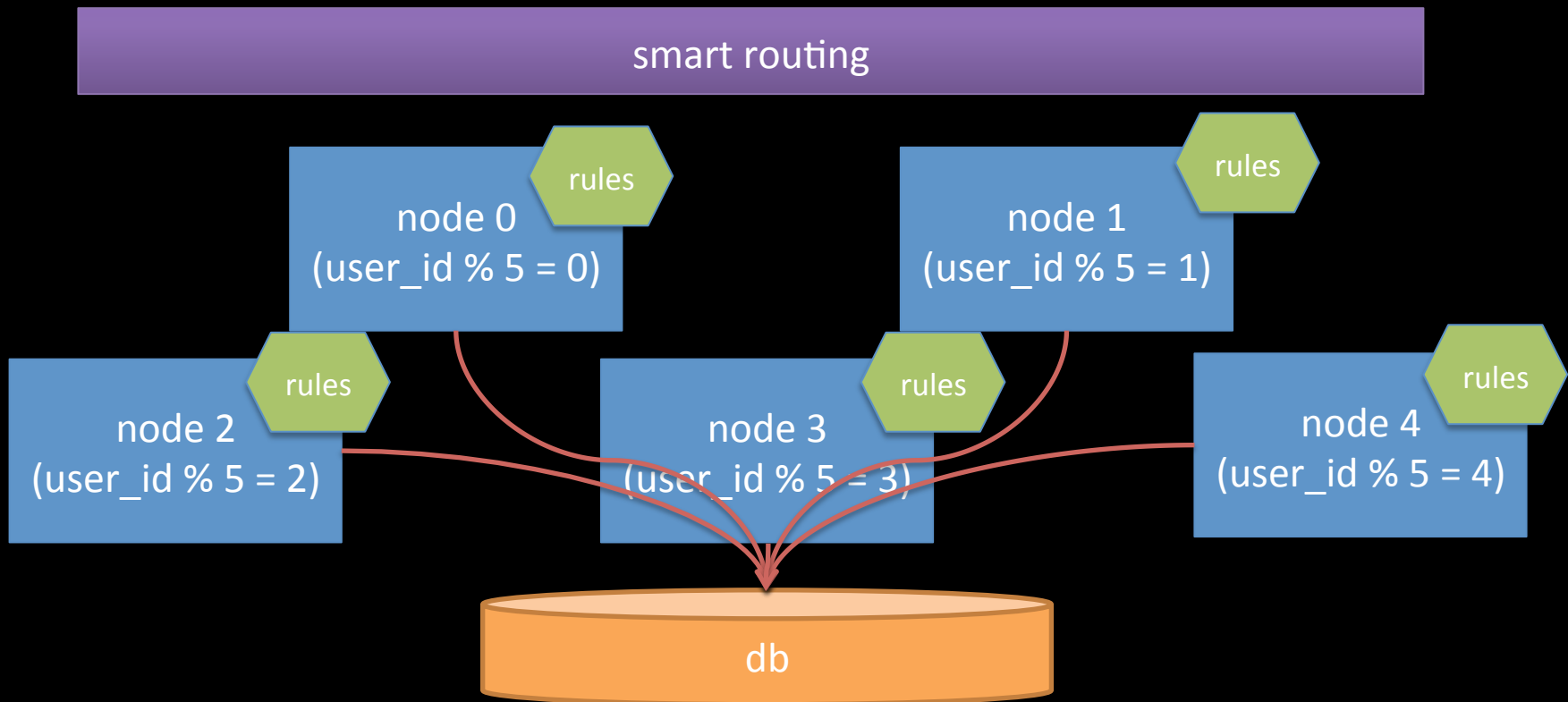
parallelization of db calls



caution: overwhelmed databases because of increased concurrent queries

get rid of db calls

- run rules with locality
- distributed hash table (dht) as data store



scalability



scalability

- volume at ebay / day
 - 30 million sign-in
 - 20 million bids/purchases
 - 15 million new listings
 - 10 million revisions
 - 8 million messages
- 15,000 fraud detection rules
- 2 rules deployments / week
- 30 million fired rules / day

horizontal scalability

- decision application hardware
 - 300 app servers for synchronous decisions
 - 160 app servers for asynchronous decisions
- decision database hardware
 - partitions
 - divide the single database into multiple ones
 - each one takes less load and contains less data

database partitions

- what's the best partition formula?
eg., a transaction table with seller, buyer, ip_address
 - data access pattern
 - find all transactions by seller-> partition by seller
 - how about find transactions by pair?
 - find all users using the same ip address -> partition by ip
 - uniformity
 - powersellers who have disproportionally large data
 - user index lookup
 1. find host by user
 2. find data by user on host

rules scalability - authoring

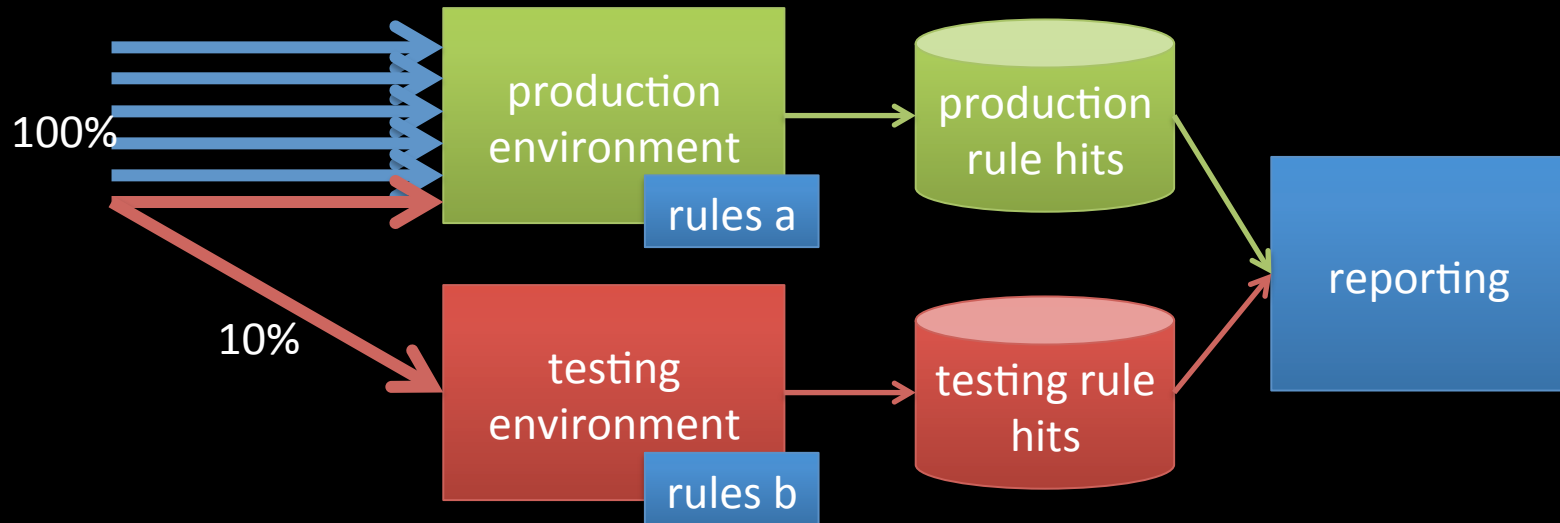
when numbers of variables, rules and analysts increase...

- management of variables
 - locate the desired variables
 - mutual correct understand the meaning of the variables
- organization of rules
 - quickly find and navigate to desired rules
- validation of rules
 - find conflicts
 - check completeness
- change control of rules
 - access control
 - approval chain
 - versioning

rules scalability - deployment

- rules deployment includes
 - validation (resources, logical constraints, etc.)
 - parsing and extraction
 - compilation
 - persistence
 - staging
- “continuous integration” of rules
 - find problems earlier and more frequently
 - deployment is ready when you need it

rules scalability – statistical testing

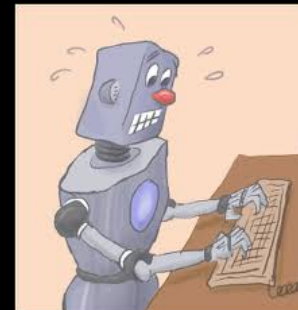


rules scalability – mark up

- attended mark up
 - hot swap of rules
 - no interruption of service
 - staging of rules, or rather, serialized objects, locally
 - phased mark up
 - new and old rules need to be able to coexist
 - validation and monitoring of rule hits
 - 1st box -> 25% -> 50% -> bake -> 100%
 - 50% of app servers == 50% expected rule hits?

rules scalability – mark up

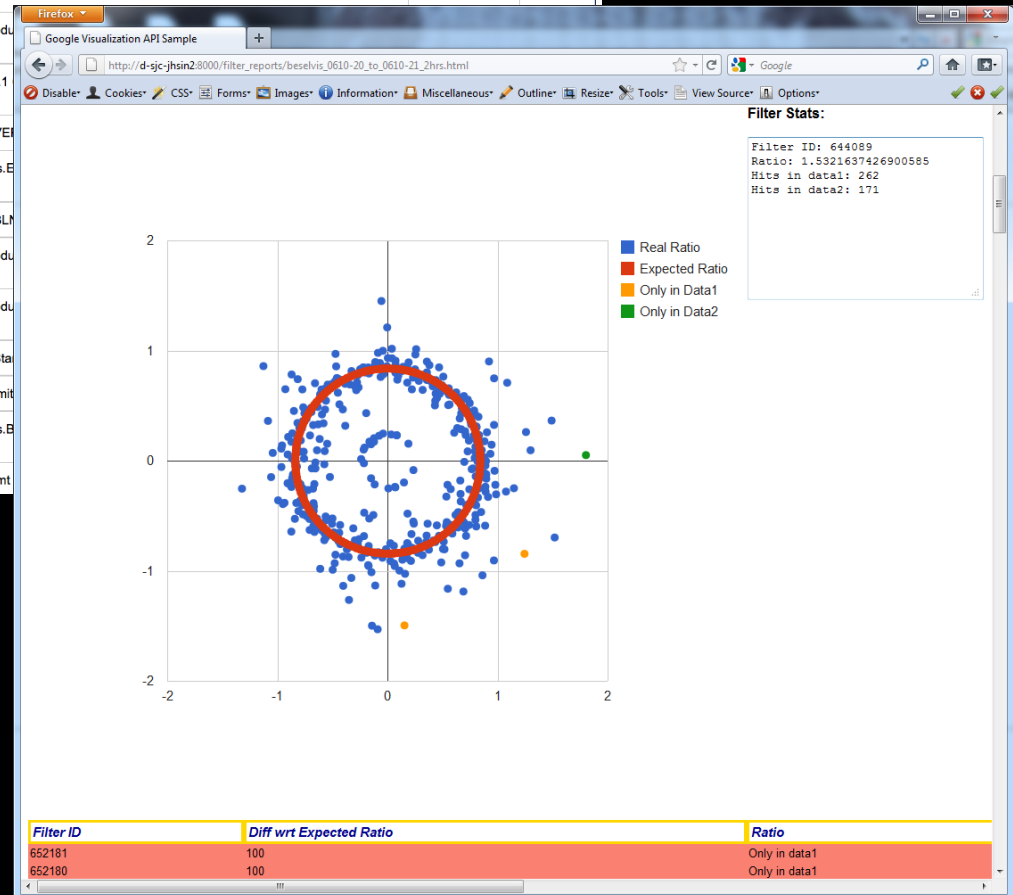
- unattended mark up
 - all the tasks from attended markup, automated
 - pause and alert administrator on exceptions
 - using rules to monitor rules
 - if volume of rule $>$ threshold and rule action = block then pause and alert
 - if volume of rule $>$ threshold and volume of rule $<$ disaster level then continue and alert
 - if volume of rule in new version $>$ 125% of old version and rule has modification then pause and alert



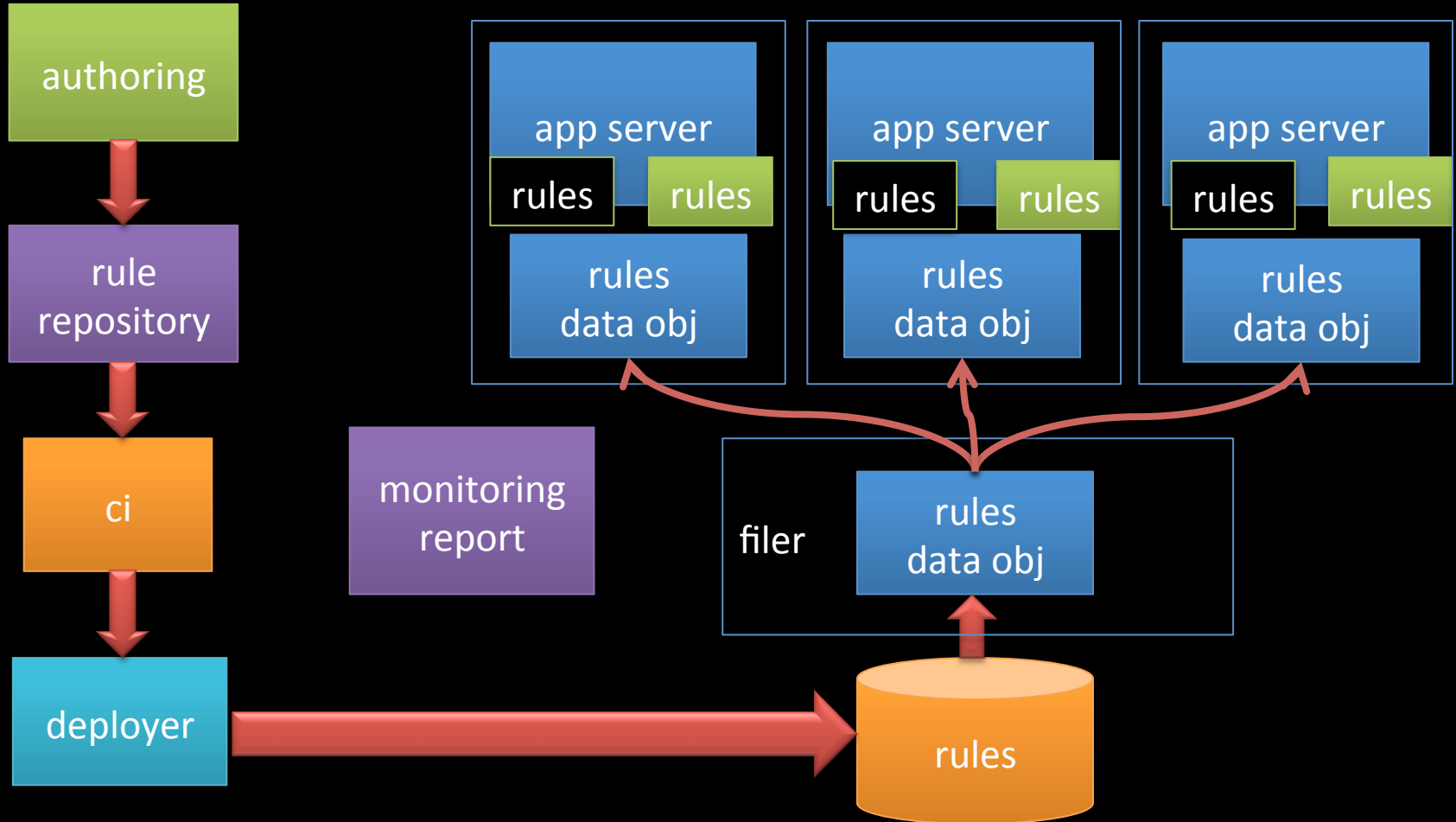
rules scalability - monitoring

Overall Top 25 Block Volume Filters

RULE_ID	RULE_AUTHOR	POLICY_GROUP_NAME	POLICY_NAME	RULE_NAME	RULE_ACTION	VOLUME
642134	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product.BSTd 642134 Below Standard Limt Exceeding Qty	Block	1036
642133	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product.BSTd 642133 Below Standard Limt Exceeding GMS	Block	418
641919	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product.SWL 641919 Exceeding GMS Limit	Block	374
643285	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product		
649596	tsehlmei	Infringement & Vero	CSA Seller Vetting	Shared.Infringement & Vero.CSA Seller Vetting.1 C2C		
168434	mpezely	Selling Limits	Global Csqn Guidelines	Shared.Selling Limits.Global Csqn Guidelines.VER		
603980	tsehlmei	Infringement & Vero	CIT Class E Blocks	Shared.Infringement & Vero.CIT Class E Blocks.E SYI		
616210	ebradford	Selling Limits	Global Csqn Guidelines	Shared.Selling Limits.Global Csqn Guidelines.BLM		
641918	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product		
642132	ebradford	Fraud & Abuse	Site Limits - Identity Product	Shared.Fraud & Abuse.Site Limits - Identity Product		
649964	cmireles	List Practices	ELG Temp	Shared.List Practices.ELG Temp.Low BIN VS Sta		
650128	tsehlmei	Fraud & Abuse	New Seller Category Limits	Shared.Fraud & Abuse.New Seller Category Limit		
595787	tsehlmei	Infringement & Vero	CIT Class B Blocks	Shared.Infringement & Vero.CIT Class B Blocks.B PID SYI		
240737	jtyler	Fraud & Abuse	UK FA SNP Pymt Restrict	UnitedKingdom.Fraud & Abuse.UK FA SNP Pymt		



rules scalability - deployments



thank you

kenny shi

kshi@ebay.com

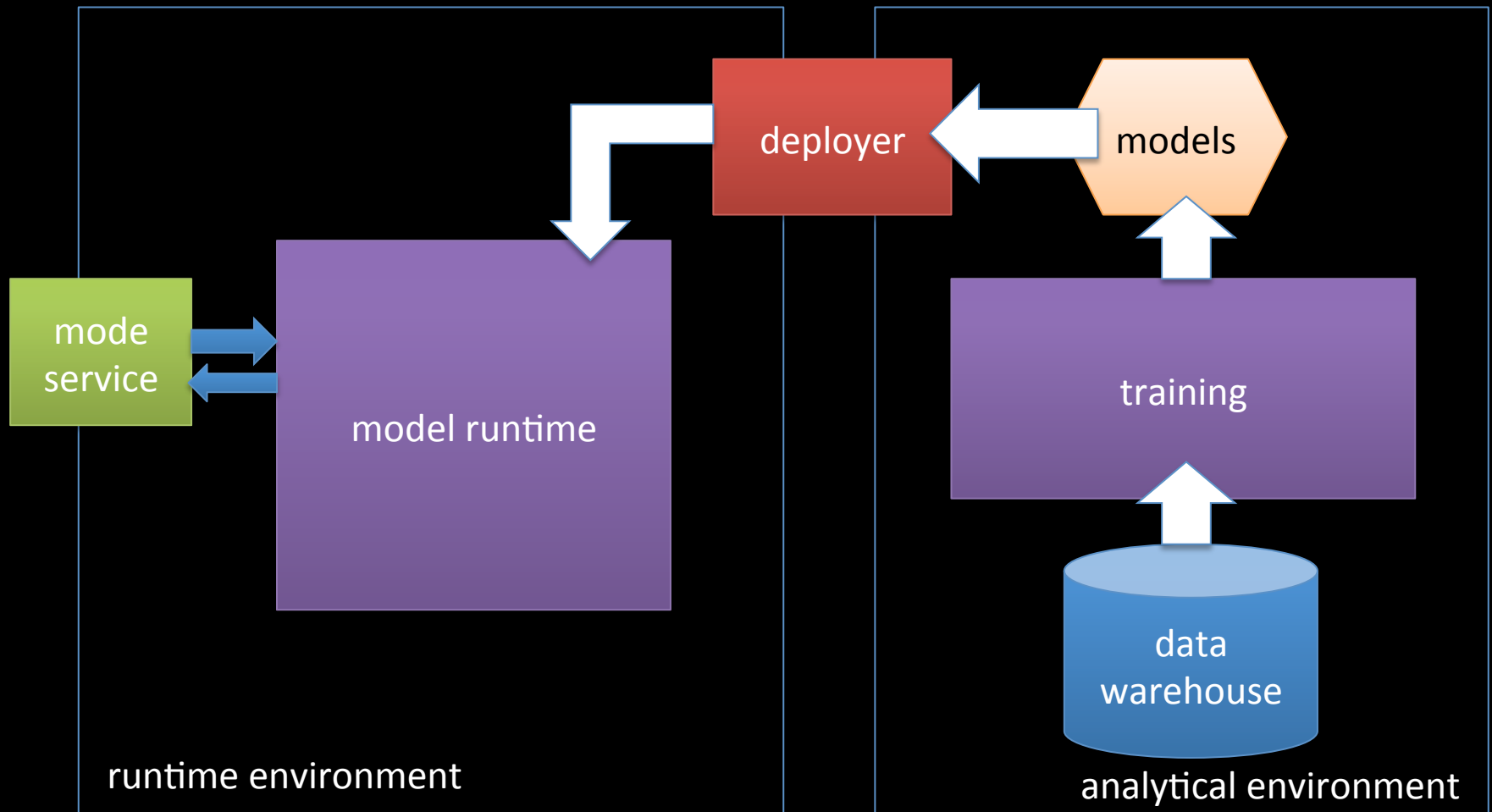
backup slides

what is a real-time decision?

- a not-so-real-time decision:
- a real-time decision:



models architecture



models deployment

- very similar approaches as rules
 - horizontally scalable runtime (65 vm internal cloud)
 - hot swappable
 - phased mark up
 - Soaking infrastructure (statistical testing)

interoperability b/w rules & models

- both rules and models need access to data
- often same data!
- rule-based data preparation
 - share abundance of rule variables
 - expressiveness of rules to populate model input
 - if user's country is usa and paypal account is on file then set value (true) for variable ("payment on file")
 - If user's country is germany and ach is on file then set value (true) for variable ("payment on file")