AEROSPIKE

NEXTGEN NOW

SUMMIT '20

Building a cashless society

PhonePe

Koushik Ramachandra

Software Architect

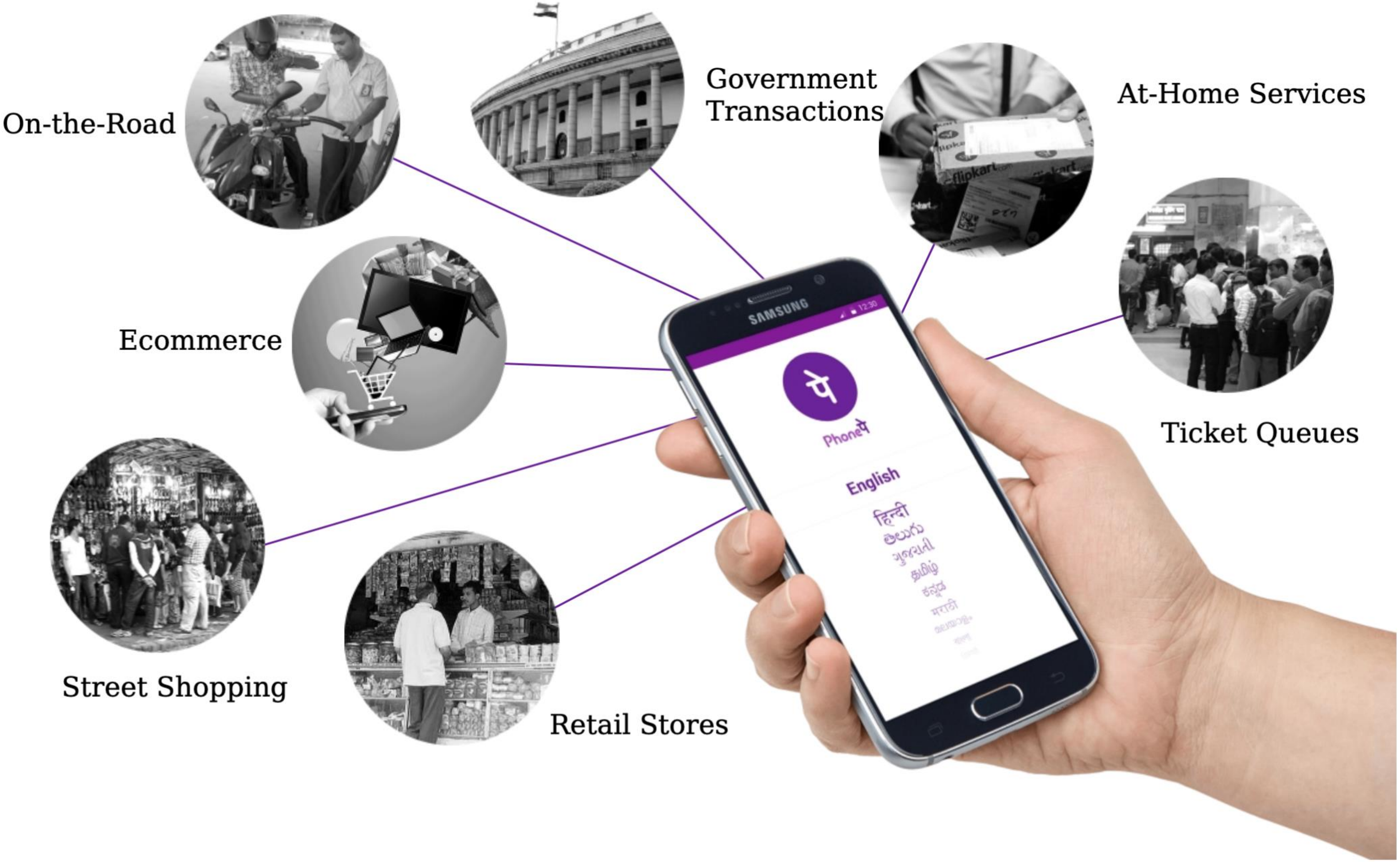
PhonePe

Agenda

- About PhonePe
- Beginning with Aerospike
- Evolution
- Varied topologies we run
- Benefits
- Offerings we use
- Learnings
- A few asks!

	People	Mobile Phone Users	Bank A/c holders	Internet Users	Smart Phone Users
	#2 in the world	#2 in the world	#2 in the world	#2 in the world	#2 in the world
2015	1,250 M	1,000 M	650 M	300 M	240 M (24%)
2020	1,350 M	1,200 M	900 M	650 M	520 M (52%)

Source: eMarketer, Ericsson, UN estimates, BCG research



On-the-Road

Government Transactions

At-Home Services

Ecommerce

Ticket Queues

Street Shopping

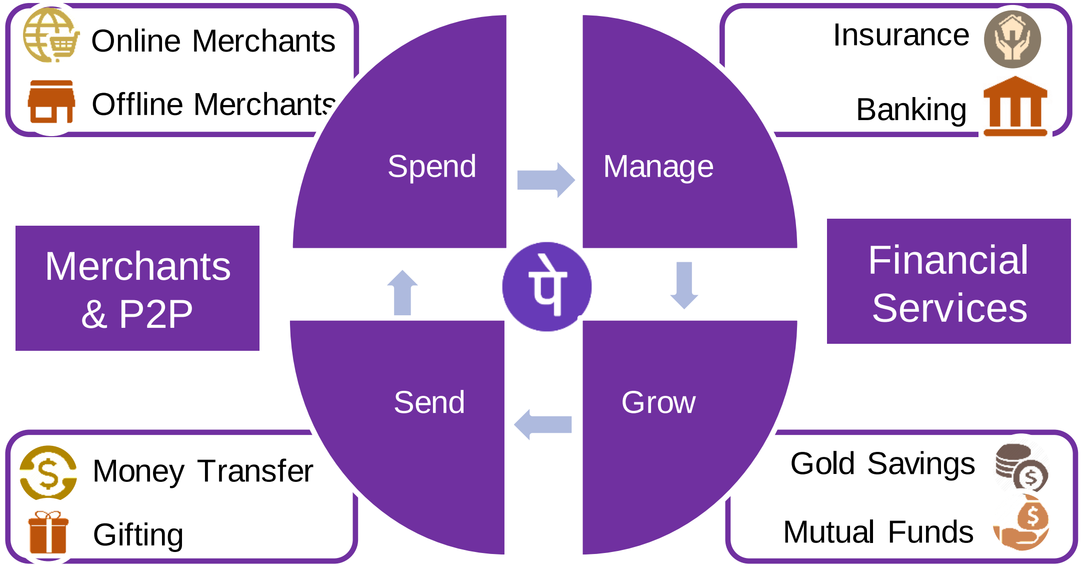
Retail Stores

Send, Spend, Manage, Grow

Partner Ecosystem Commerce

- Retail Stores
- Kirana Stores
- Online Commerce
- Travel and Commute
- Food
- Education
- Healthcare

PhonePe Platform Consumer Journey



Online Merchants
Offline Merchants
Spend
Manage
Insurance
Banking
Merchants & P2P
Financial Services
Send
Grow
Money Transfer
Gifting
Gold Savings
Mutual Funds

Partner Ecosystem Financial Services

- Insurance
- Banking
- Credit
- Gold Savings
- Mutual Funds

What we do?

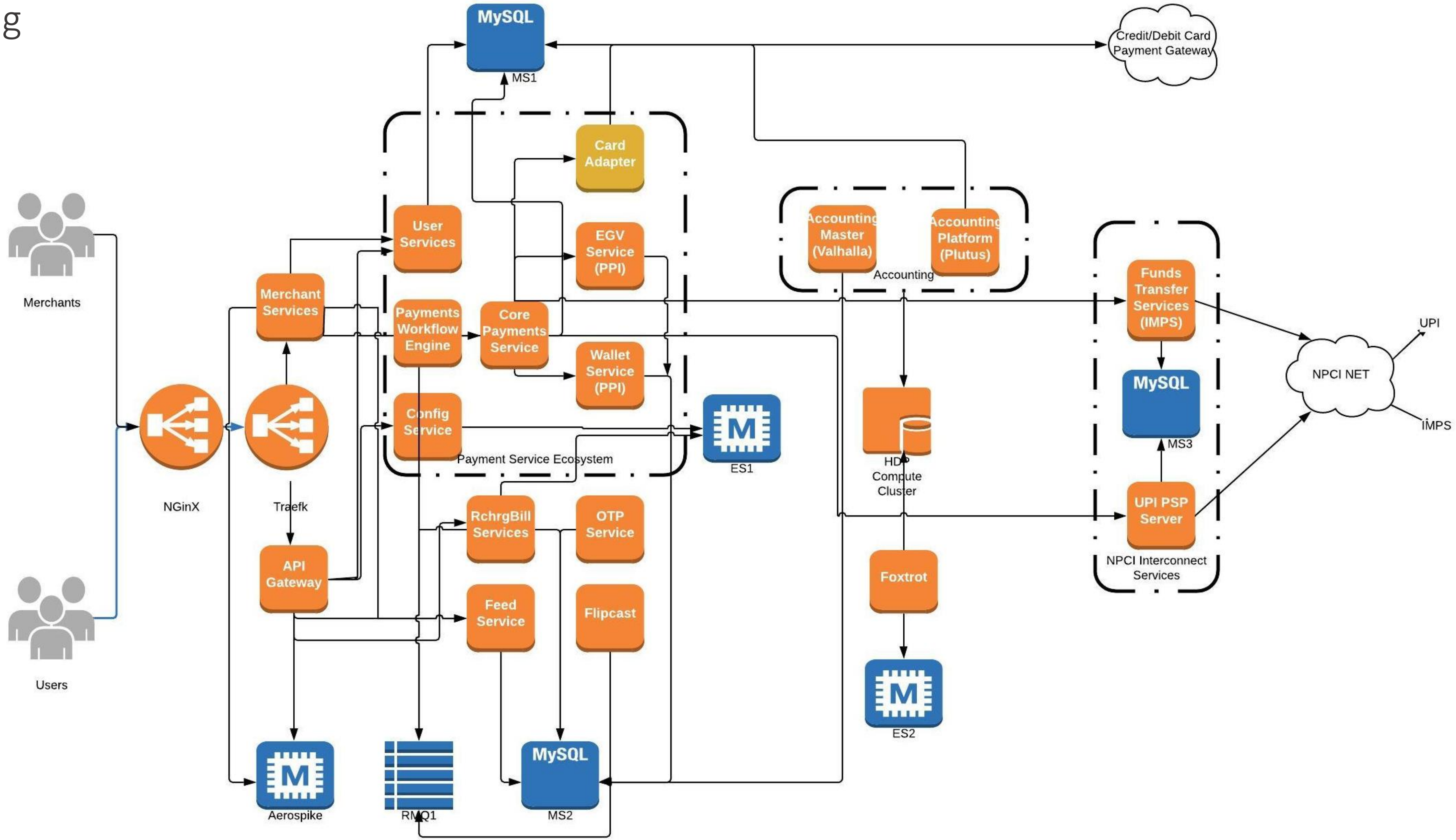
- 200M users
- 80M MAU
- ATV - $180+ bn
- Merchants - 10M across 400 cities
- 500M+ monthly transactions

Flipkart
JABONG COM
ekart LOGISTICS
MYNTRA.com
McDonald's
zomato
FreshMenu
TICKETNEW
Café Coffee Day a LOT can happen over coffee
cleartrip
Mumbai Metro
METRO Cash & Carry
Apollo PHARMACY
BARISTA
bookmyshow
IndianOil
!dea
SWIGGY FOOD DELIVERY APP
abhibus
KFC
indiamart
OLA
spencer's Makes fine living affordable
Faasos
redBus.in
make my trip
IRCTC
PVR CINEMAS

Guiding Principles

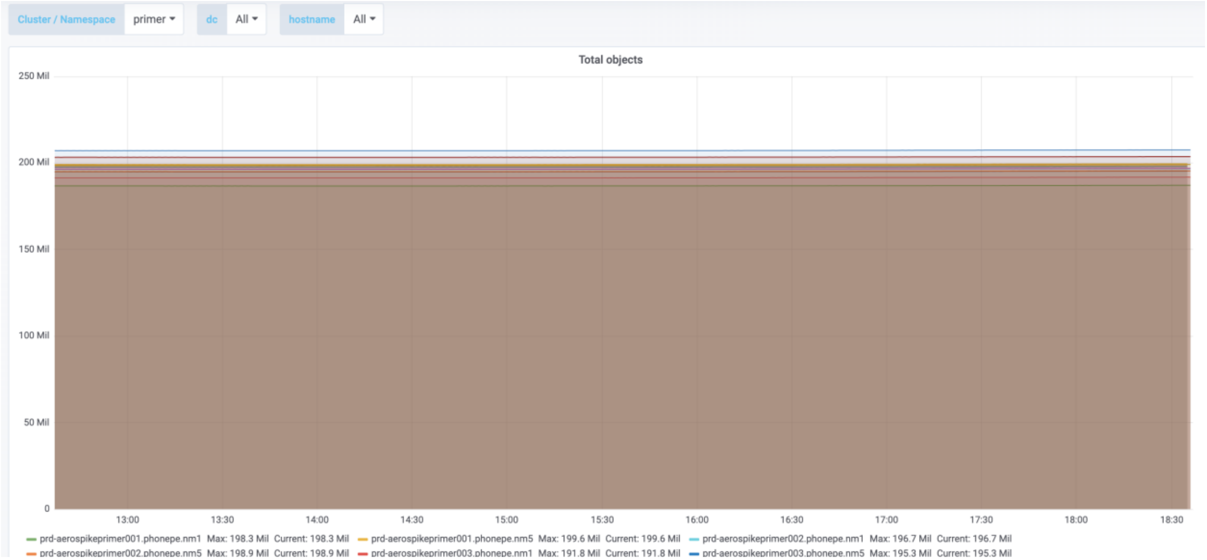
- Ability to handle huge volume of traffic
- Easily grow with traffic
- Resource utilisation
- Fair work distribution
- Monitoring
- Isolation & Shared-nothing
- Resilience
- XDR
- Capacity Planning

The beginning

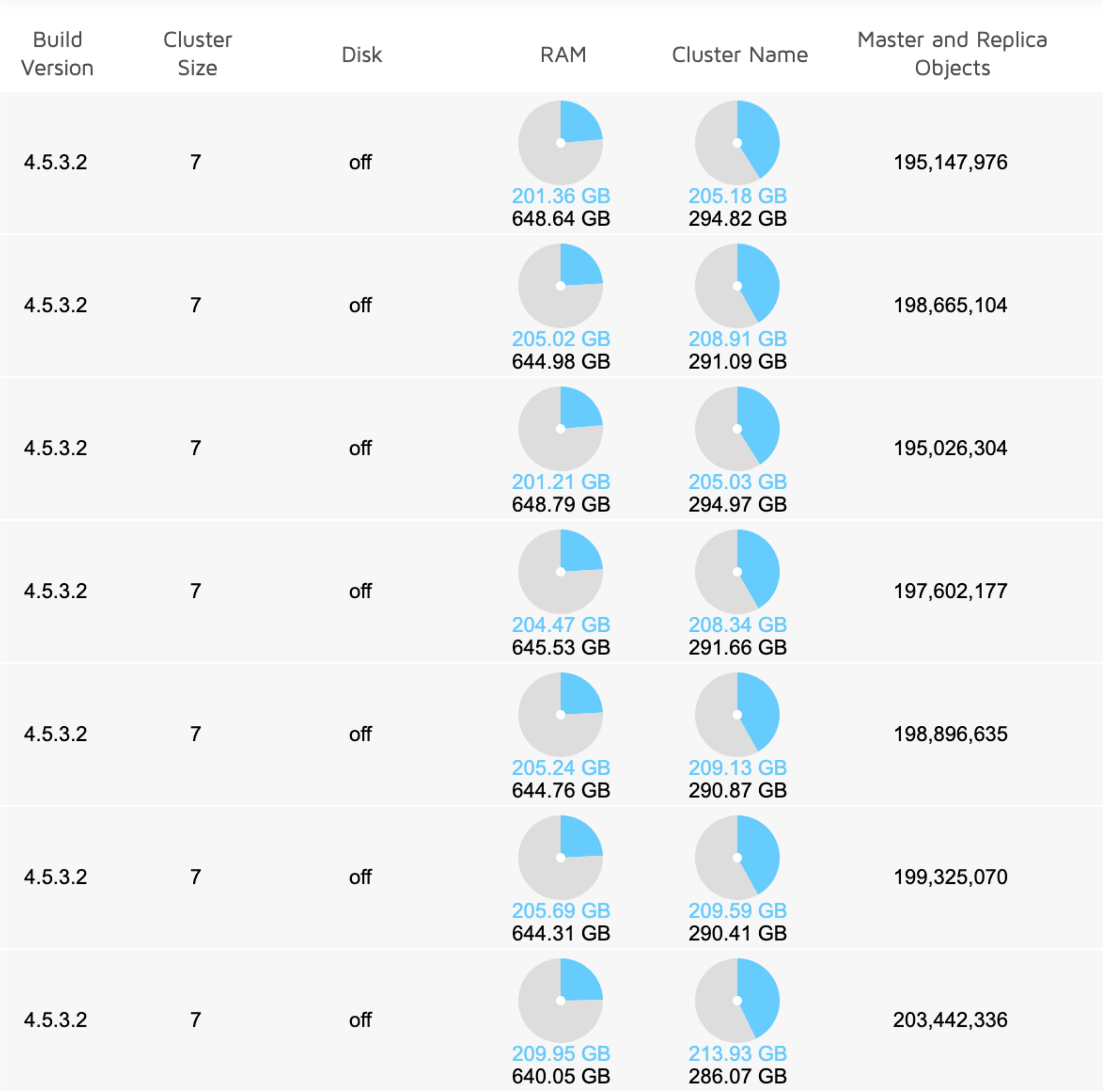


Mailbox & Primer - Edge

- JWT Authorization
- Mobile Scale Request Polling
- 200K QPS
- < 1 ms
- > 200M objects
- Data in memory



Cluster / Namespace: primer
dc: All
hostname: All



Build Version	Cluster Size	Disk	RAM	Cluster Name	Master and Replica Objects
4.5.3.2	7	off	201.36 GB 648.64 GB	205.18 GB 294.82 GB	195,147,976
4.5.3.2	7	off	205.02 GB 644.98 GB	208.91 GB 291.09 GB	198,665,104
4.5.3.2	7	off	201.21 GB 648.79 GB	205.03 GB 294.97 GB	195,026,304
4.5.3.2	7	off	204.47 GB 645.53 GB	208.34 GB 291.66 GB	197,602,177
4.5.3.2	7	off	205.24 GB 644.76 GB	209.13 GB 290.87 GB	198,896,635
4.5.3.2	7	off	205.69 GB 644.31 GB	209.59 GB 290.41 GB	199,325,070
4.5.3.2	7	off	209.95 GB 640.05 GB	213.93 GB 286.07 GB	203,442,336

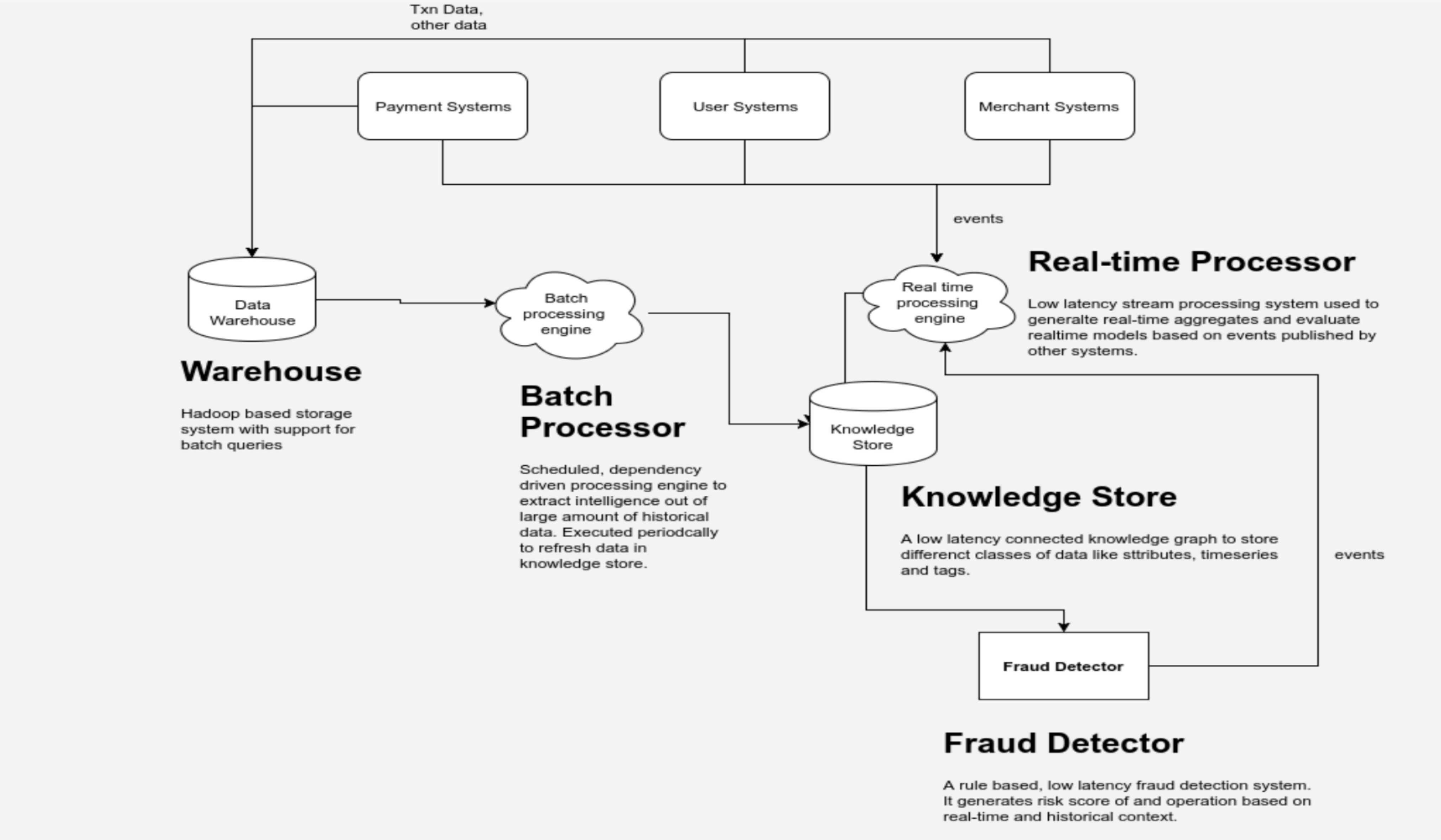
We run

- 30+ clusters of varying sizes
- Heterogenous profiles
- Billions of records
- < 3 ms 99.9
- XDR enabled
- Not just speed (Data complexity)
- Consistency

How?

- Microservices
- Isolation per service
- Profile independently
- Give developers a bit of freedom and evolve
- Treat SSDs as cheap RAM not pricey disk
- Sizing

Compute



Real-time Processor

Low latency stream processing system used to generalte real-time aggregates and evaluate realtime models based on events published by other systems.

Warehouse

Hadoop based storage system with support for batch queries

Batch Processor

Scheduled, dependency driven processing engine to extract intelligence out of large amount of historical data. Executed periodcally to refresh data in knowledge store.

Knowledge Store

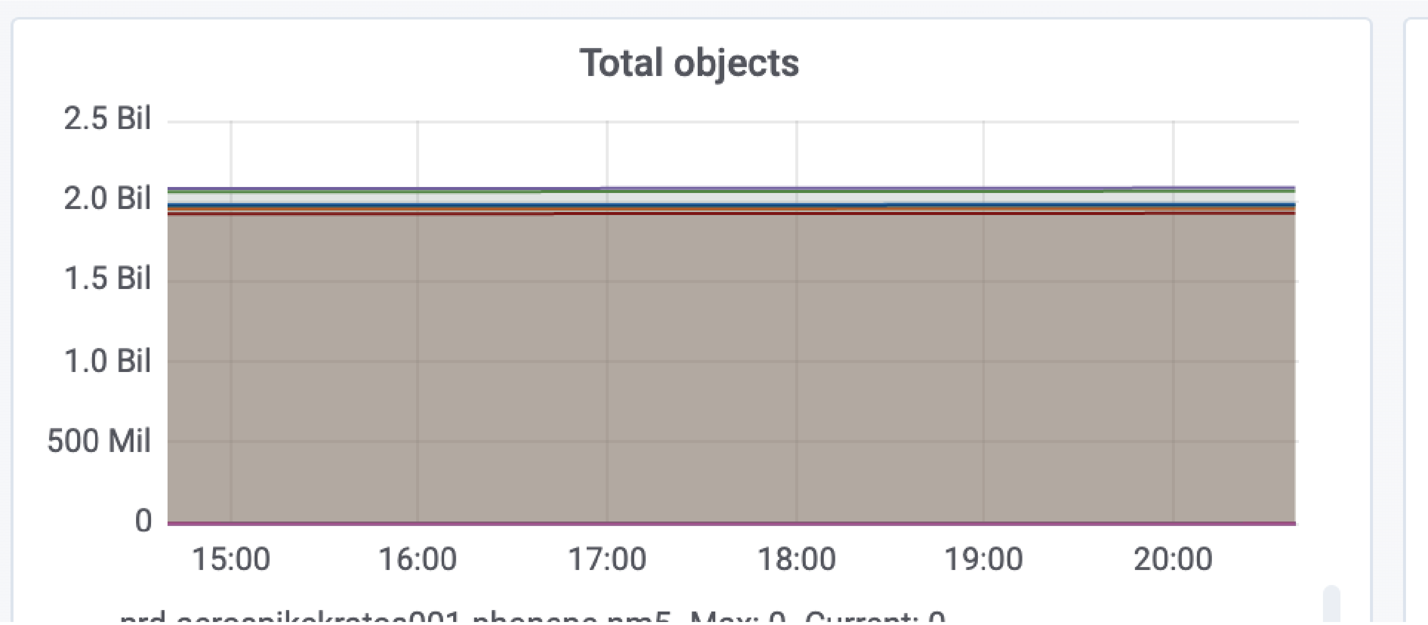
A low latency connected knowledge graph to store differenct classes of data like sttributes, timeseries and tags.

Fraud Detector

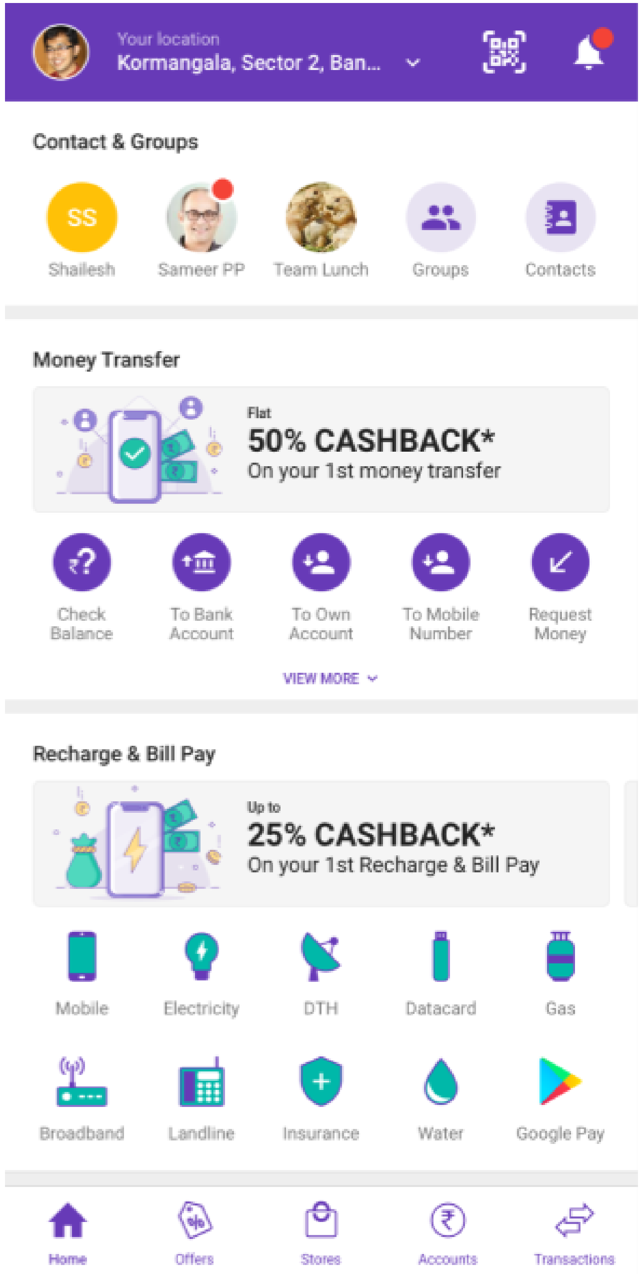
A rule based, low latency fraud detection system. It generates risk score of and operation based on real-time and historical context.

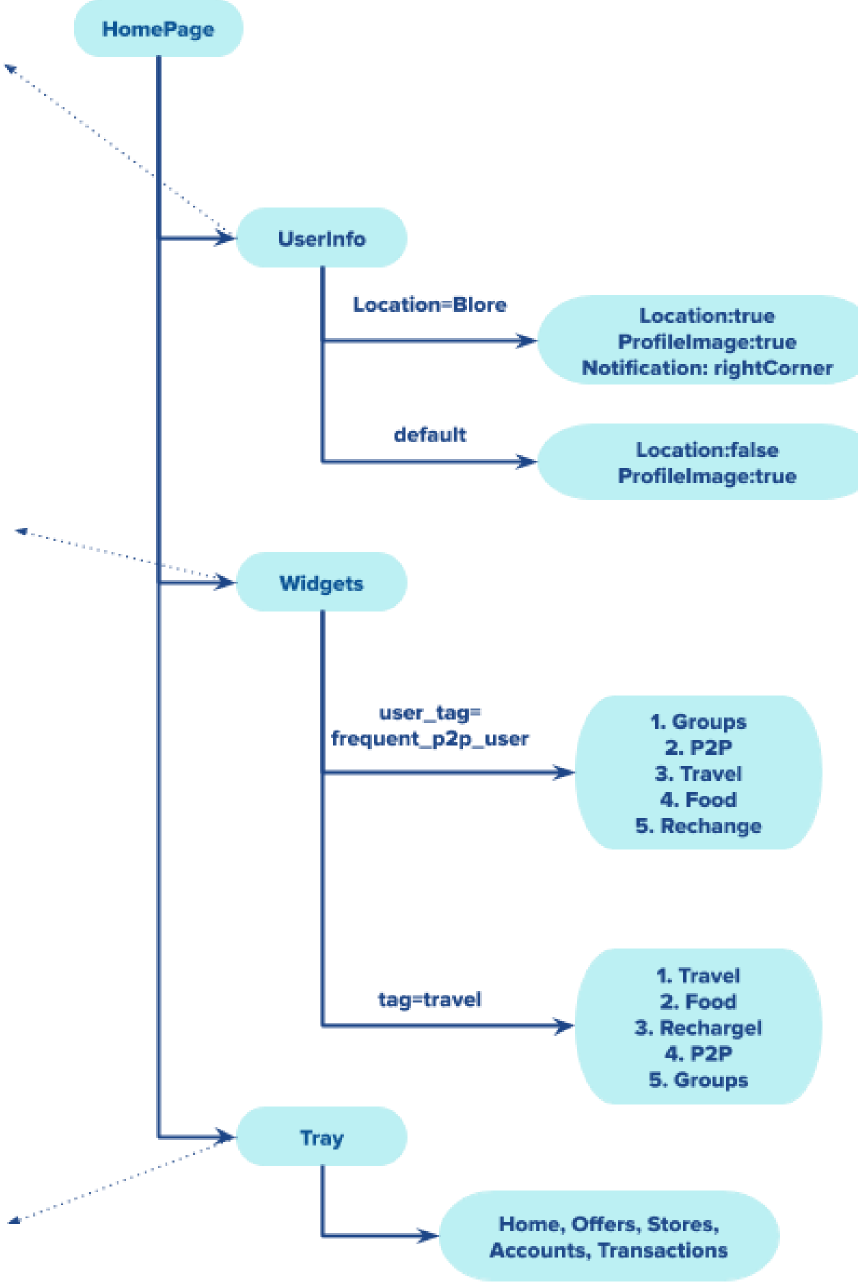
Compute Scale

- Growth, Fraud
- Inline
- Mixed workload
- Complex data types
- ~ 2B objects
- ~ 200K QPS
- < 2 ms 99th - multi read
- Storage: Disk



Personalization





HomePage

UserInfo

Location=Blore → Location:true, ProfileImage:true, Notification: rightCorner

default → Location:false, ProfileImage:true

Widgets

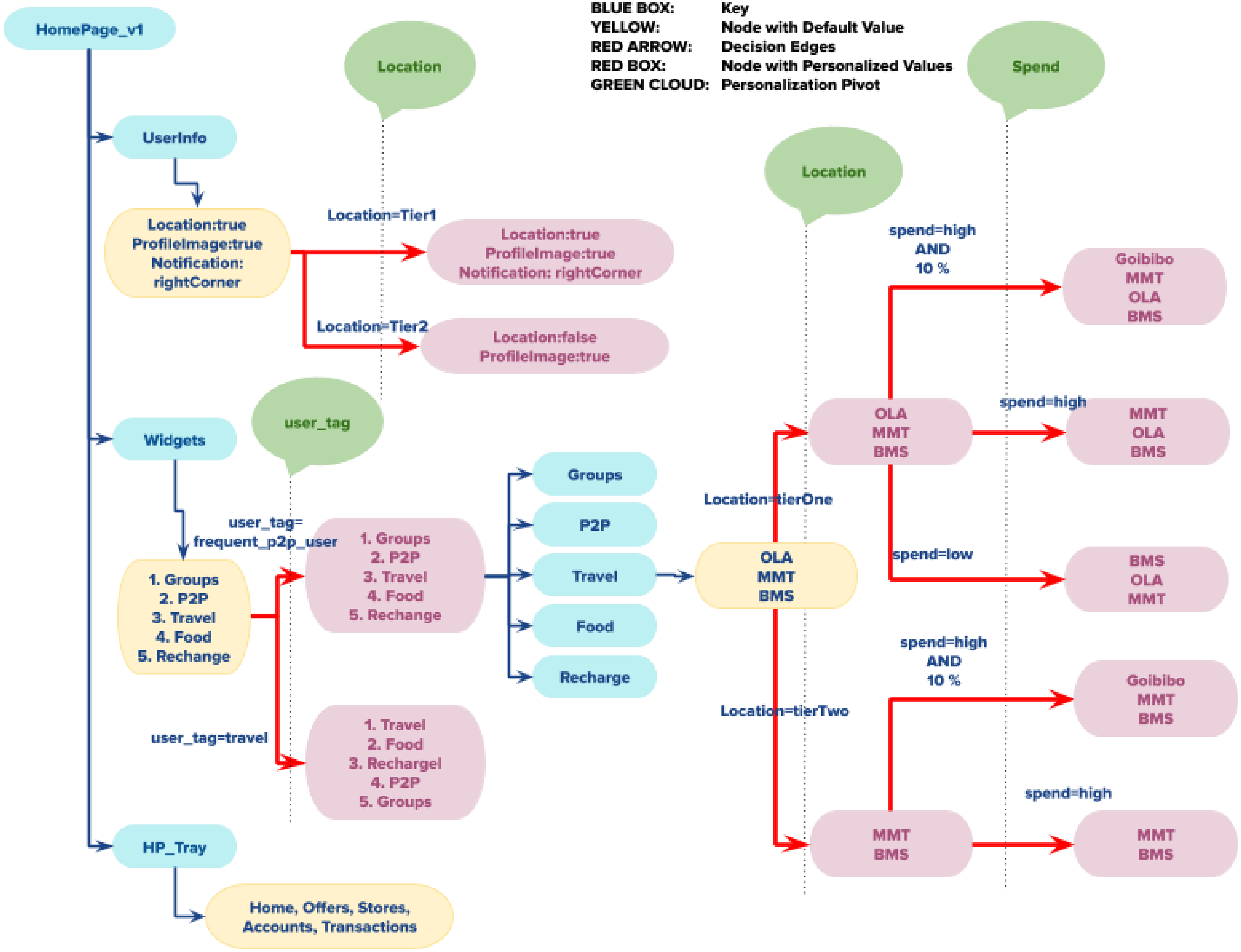
user_tag= frequent_p2p_user → 1. Groups 2. P2P 3. Travel 4. Food 5. Rechange

tag=travel → 1. Travel 2. Food 3. Rechargel 4. P2P 5. Groups

Tray

Home, Offers, Stores, Accounts, Transactions

BLUE BOX:	Key
YELLOW:	Node with Default Value
RED ARROW:	Decision Edges
RED BOX:	Node with Personalized Values
GREEN CLOUD:	Personalization Pivot



HomePage_v1

UserInfo

Location:true
ProfileImage:true
Notification:
rightCorner

Location

Location=Tier1

Location:true
ProfileImage:true
Notification: rightCorner

Location=Tier2

Location:false
ProfileImage:true

Widgets

user_tag

user_tag=
frequent_p2p_user

1. Groups
2. P2P
3. Travel
4. Food
5. Rechange

1. Groups
2. P2P
3. Travel
4. Food
5. Rechange

user_tag=travel

1. Travel
2. Food
3. Rechargel
4. P2P
5. Groups

Groups

P2P

Travel

Food

Recharge

OLA
MMT
BMS

Location

Location=tierOne

OLA
MMT
BMS

Location=tierTwo

MMT
BMS

Spend

spend=high
AND
10 %

Goibibo
MMT
OLA
BMS

spend=high

MMT
OLA
BMS

spend=low

BMS
OLA
MMT

spend=high
AND
10 %

Goibibo
MMT
BMS

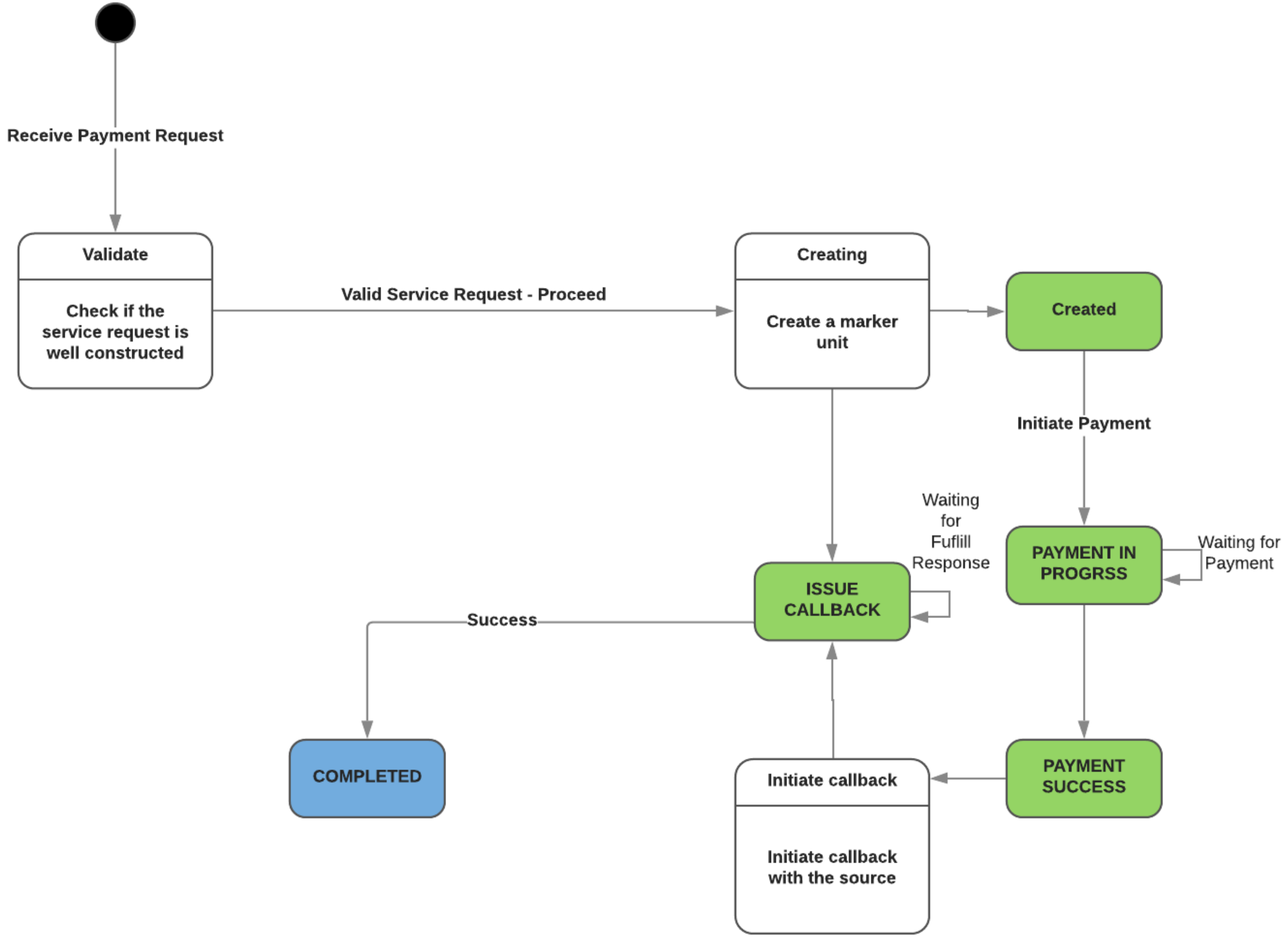
spend=high

MMT
BMS

HP_Tray

Home, Offers, Stores,
Accounts, Transactions

From a traditional pattern…



Classical data models

TransactionId	Amount	merchantId	providerId
TRAFIMY#31$	200000	M1	P1
TRAFIMY$212	121131	M2	P2

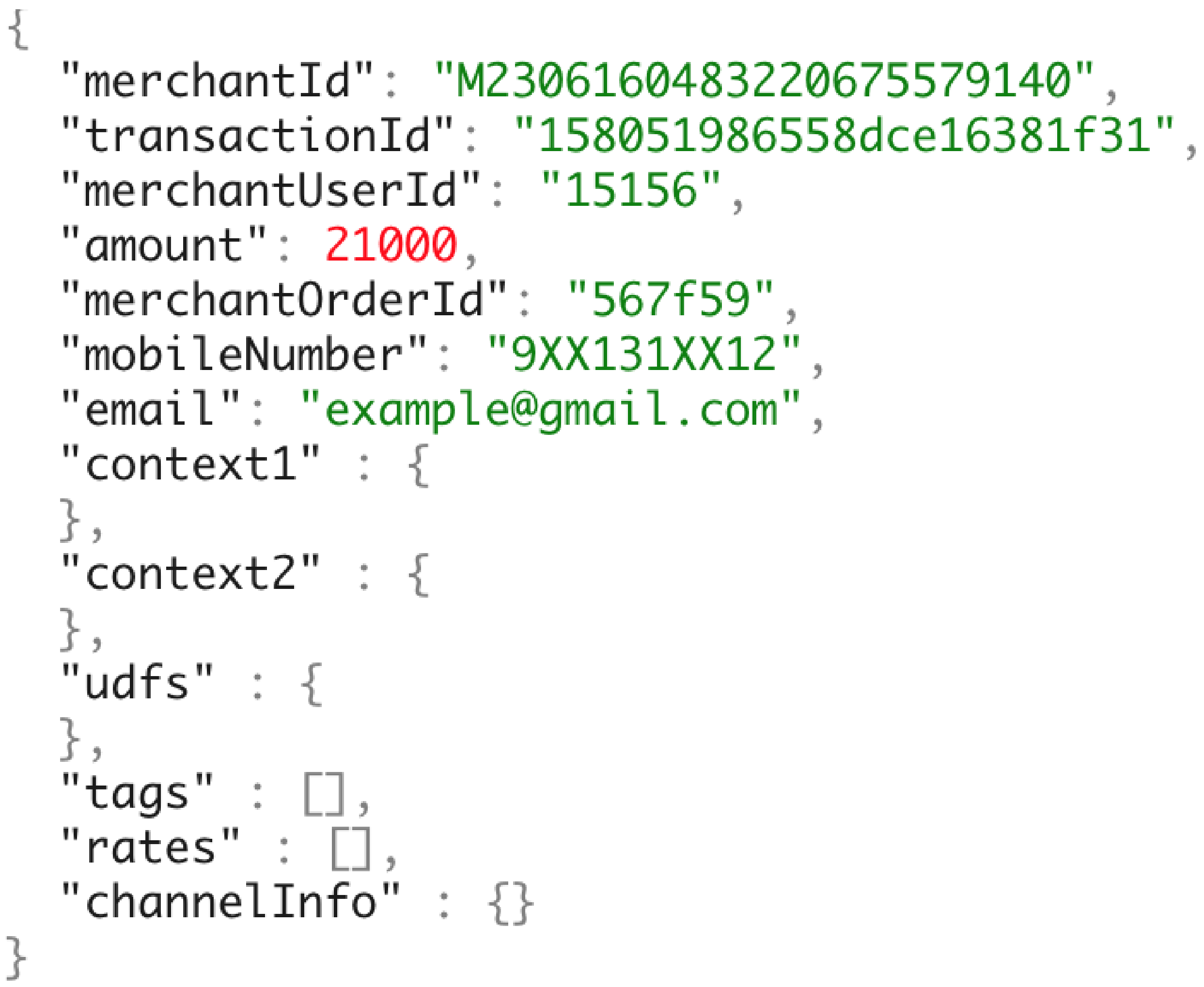
Growth? Shard?

TransactionId	Amount	merchantId	providerId	PartitionId
TRAFIMY#31$	200000	M1	P1	1
TRAFIMY$212	121131	M2	P2	2

Archive? TTL?

How?

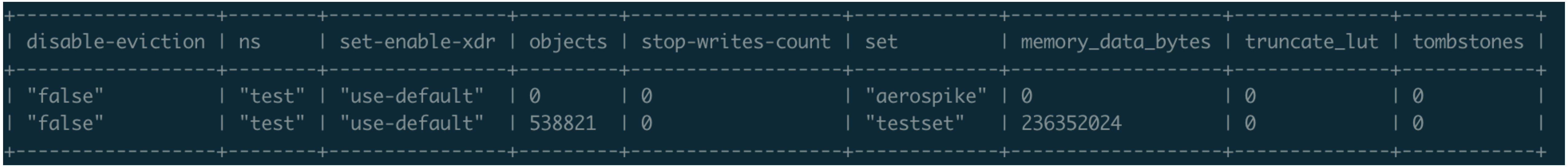
- Atomic
- Idempotence
- Faceting
- Evolution of proxies
- New World Imaging
- Storage : Disk
- < 2 ms 99th - multi read
- Security
- Maintainability
- Getup fast!



```
{
  "merchantId": "M2306160483220675579140",
  "transactionId": "15805198655 8dce16381f31",
  "merchantUserId": "15156",
  "amount": 21000,
  "merchantOrderId": "567f59",
  "mobileNumber": "9XX131XX12",
  "email": "example@gmail.com",
  "context1" : {
  },
  "context2" : {
  },
  "udfs" : {
  },
  "tags" : [],
  "rates" : [],
  "channelInfo" : {}
}
```

Spatial Data

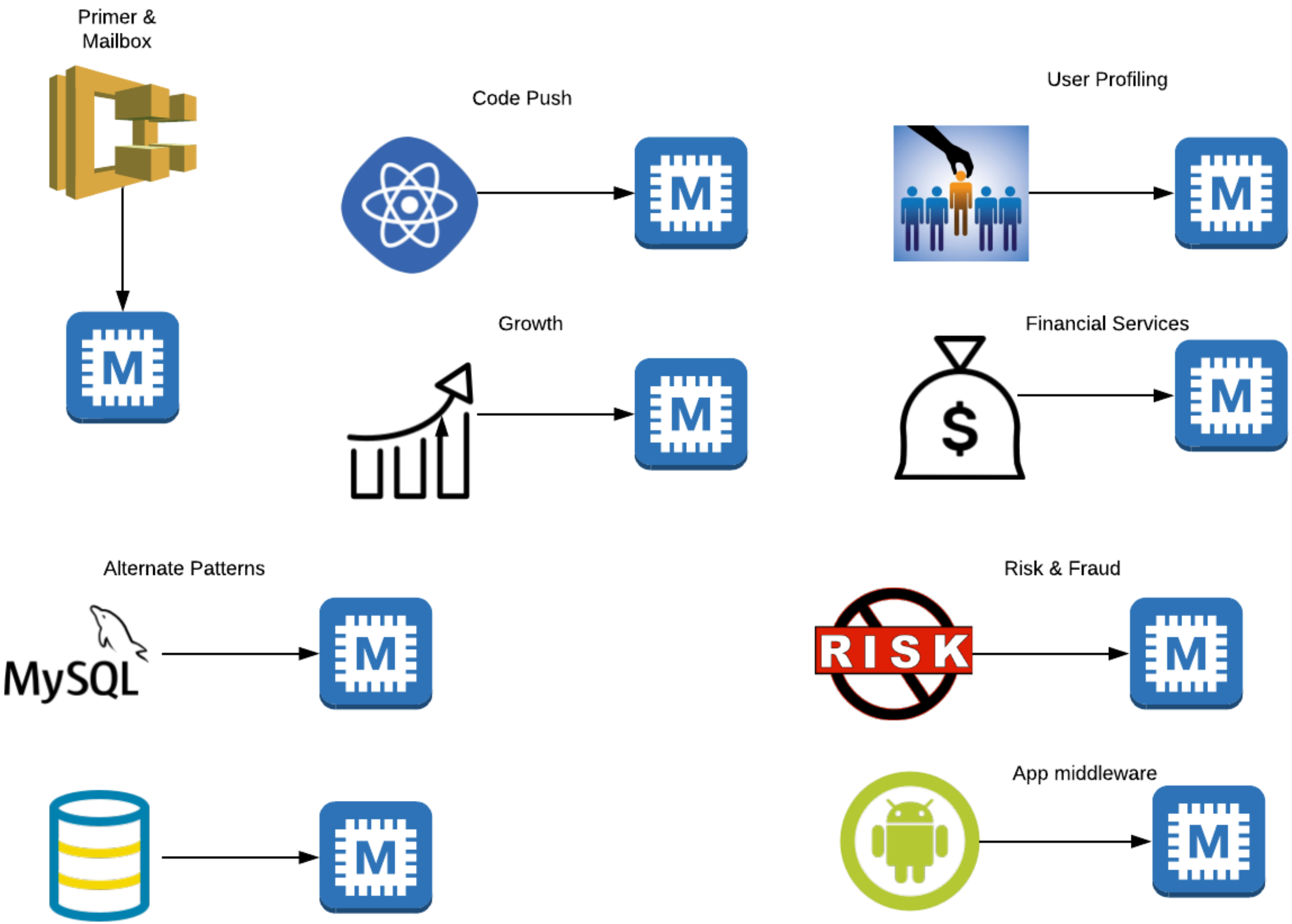
- Location tagged to all data
- Fast near-by searches
- Enter/Exit geofences
- Region containing point



```
+------------------+--------+----------------+---------+-------------------+-------------+-------------------+--------------+------------+
| disable-eviction | ns     | set-enable-xdr | objects | stop-writes-count | set         | memory_data_bytes | truncate_lut | tombstones |
+------------------+--------+----------------+---------+-------------------+-------------+-------------------+--------------+------------+
| "false"          | "test" | "use-default"  | 0       | 0                 | "aerospike" | 0                 | 0            | 0          |
| "false"          | "test" | "use-default"  | 538821  | 0                 | "testset"   | 236352024         | 0            | 0          |
+------------------+--------+----------------+---------+-------------------+-------------+-------------------+--------------+------------+
```

- 16 vertices covering India
- 50K queries, result set length = 10, concurrency = 8
- P99 -> 2.2s, mean -> 444 ms
- Varied TPS! - Density.
- Modelling ain’t easy as well.

Evolution : “Nearly” all things to everyone!



Monitoring

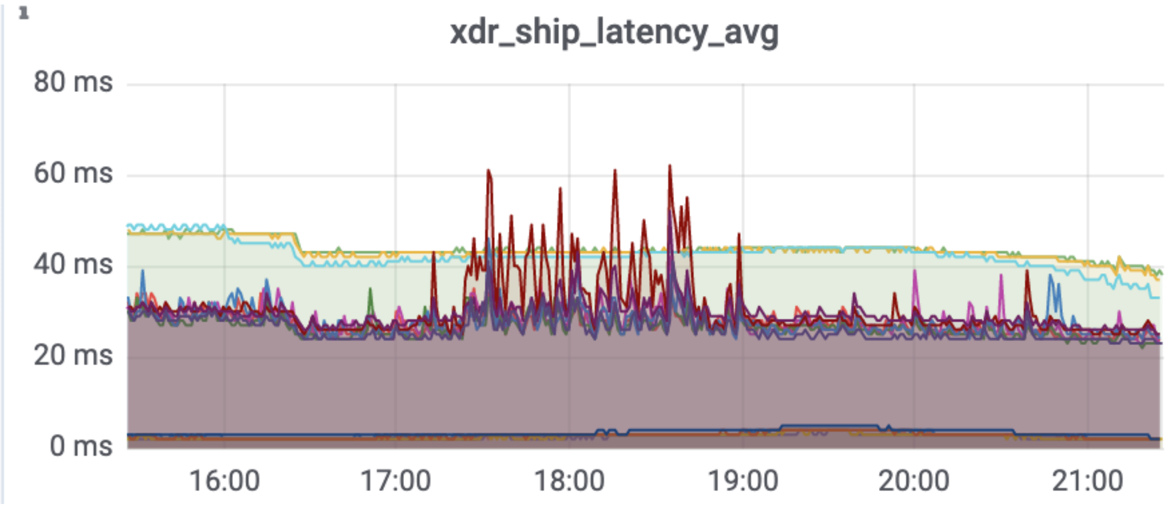
- DC aware
- One stop shop
- Riemann/InfluxDB/Grafana
- For a sample metric collector, check this out

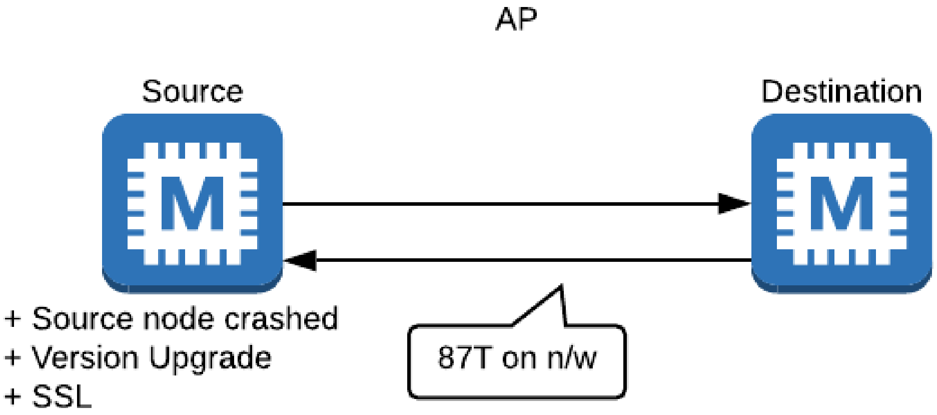
Benefits

- Off the cuff
- New World Imaging
- Capacity Planning
- Shared nothing
- Predictability
- Cost
- Storage Models
- Scalability
- Monitoring

XDR : AP

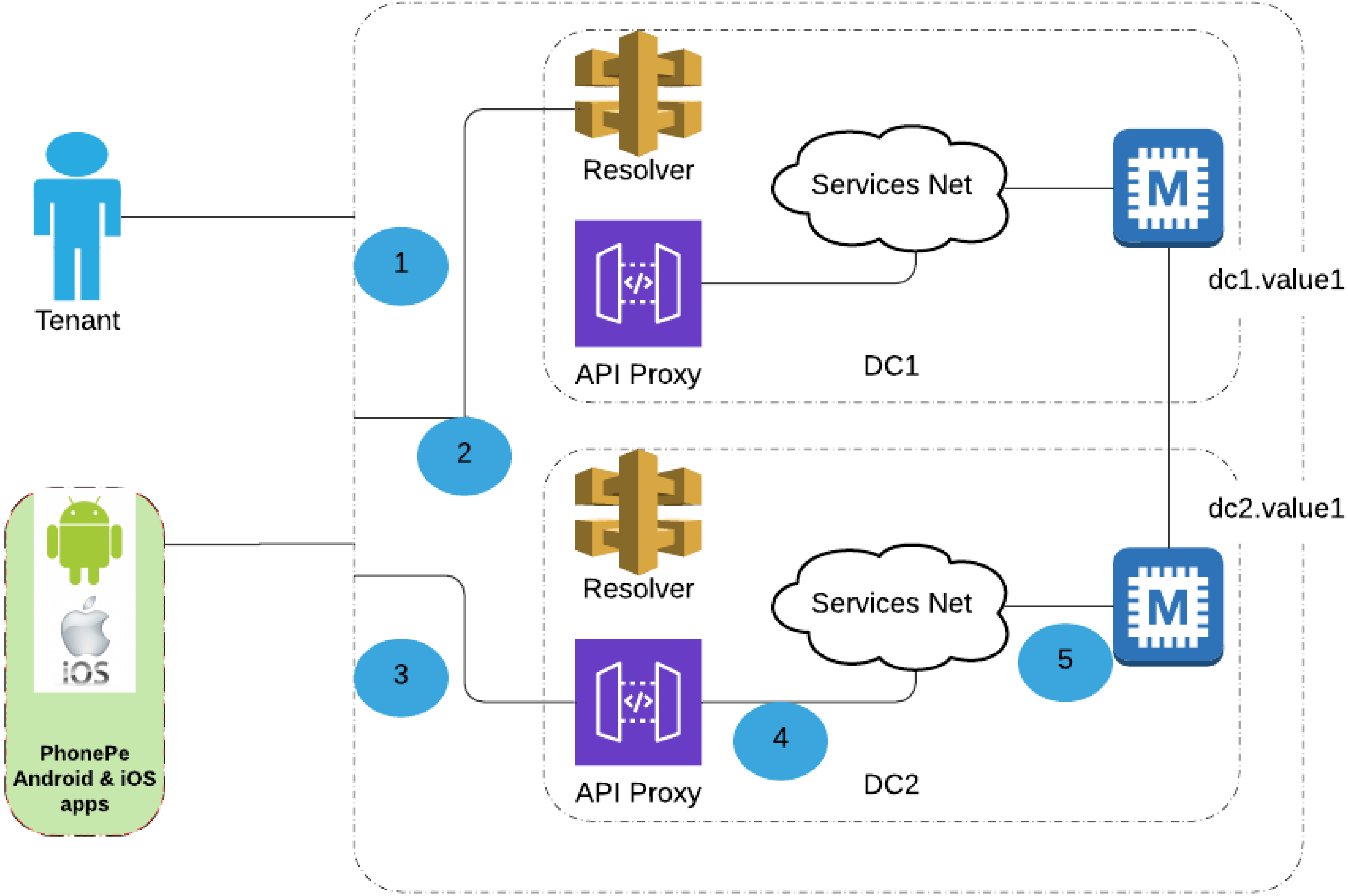
- As slow as the slowest DC - Hey 5.0! (Add-on)
- Migrations
- Characteristics under clock/heartbeats
- Checksumming





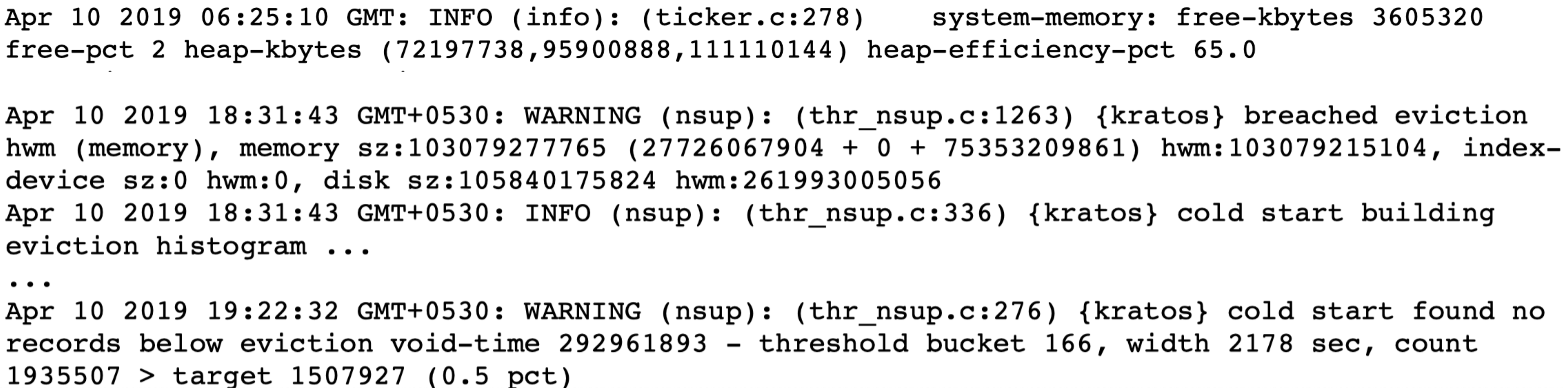
+ Source node crashed
+ Version Upgrade
+ SSL

XDR - AA



- Active-Active - Concurrent updates
- WAN latencies
- Application heavy-lifting - Faceting
- Aggregates?

Memory Allocation - OOM



```
Apr 10 2019 06:25:10 GMT: INFO (info): (ticker.c:278)    system-memory: free-kbytes 3605320
free-pct 2 heap-kbytes (72197738,95900888,111110144) heap-efficiency-pct 65.0

Apr 10 2019 18:31:43 GMT+0530: WARNING (nsup): (thr_nsup.c:1263) {kratos} breached eviction
hwm (memory), memory sz:103079277765 (27726067904 + 0 + 75353209861) hwm:103079215104, index-
device sz:0 hwm:0, disk sz:105840175824 hwm:261993005056
Apr 10 2019 18:31:43 GMT+0530: INFO (nsup): (thr_nsup.c:336) {kratos} cold start building
eviction histogram ...
...
Apr 10 2019 19:22:32 GMT+0530: WARNING (nsup): (thr_nsup.c:276) {kratos} cold start found no
records below eviction void-time 292961893 - threshold bucket 166, width 2178 sec, count
1935507 > target 1507927 (0.5 pct)
```

- Cold-start eviction, OOM during eviction
- data-in-memory=true
- 60% HWM is no blanket
- Capacity Planning
- 0% available space!

used_m_pct + [used_mem_pct / (# of nodes - accepted_#_of_failure_nodes)] + 3 %

Backup & Restore

- Tune like that suits you! - Be wary of the scans though!
- Restore & Loader utility! - Use native client library instead
- Try backing up namespaces - Avoid multiple scans
- The gzip hurdle!
- Backup into formats
- Save time compression? (single-bin - application compression is the alternate)
- Incremental backups

NUMA Pinning

- Locality of Reference
- auto-pin
- Recommendation : Don’t!
    - Non-trivial
    - Sizing becomes a hassle
    - Maintenance

And…

- asvalidation helps, but for detection
- —cdt-fix-ordered-list-unique - has version compatibility
- CDTs may need application level heavy lifting
- XDR may slow down during migrations
- Governance
- Container Environments
- Strong Consistency
- Quotas
- Probabilistic Data Structures
- On OSX

And…

- AerospikeClient for both sync and async operations
- LoggingFacility
- WritePolicy.sendKey
- Don’t use the Value or Bin constructors
- recordExistsAction : REPLACE policy. UPDATE does a read ops from disk
- sleepBetweenRetries
- Batch reads, record too big - the dreaded 22!

Summary

- It just works, happy people!
- Excellent support
- Smaller teams, no maintenance headache
- Predictability
- Heterogenous business use-cases
- Scales without hassle to millions

Thank you