



# Relevance & Personalization with Real Time Big Data Analytics

**GROUPON**<sup>®</sup>

Ameya Kanitkar  
ameya@groupon.com

## About Me

- Lead Engineer working on Real Time Data Infrastructure @ Groupon
- Graduate of Carnegie Mellon University & Pune University

# What are Groupon Deals?

**GROUPON** Featured Deal All Deals Now! Deals Getaways Sign Up Stan W. ▾

San Francisco ▾ Get Deals By Email

**Refer Friends, Get \$10**

---

**Live2kite – Greenbrae**  
Two-Hour Kiteboarding Lesson on Land or Three-Hour Kiteboarding Lesson in Water

from **\$95** **Buy!**

Value	Discount	You Save
\$200	53%	\$105

**Buy it for a friend!**

**Time Left To Buy**  
1 day 23:51:46

**Over 63 bought**  
Limited quantity available

**The deal is on!**  
Tipped at 12:40PM with 10 bought

**In a Nutshell**  
Experienced instructors guide students of every skill level as they learn to harness wind on land or in water.

**The Fine Print**  
Expires Mar 27, 2012  
Limit 1 per person, may buy 1 additional as a gift. Limit 1 per visit. Valid only for option purchased. Must be 9 or older. Must sign waiver; under 18 must have guardian sign waiver. Subject to weather. Semiprivate lesson in water valid only for those who have completed introductory lesson on land. Lessons are non-transferable. See the rules that apply to all deals.

**How far is this from home?** Add Home

**Adrenaline** **Once in a Lifetime** **Fresh Air** Get more of what you love

---

The wind has entertained humans for millennia, lofting paper airplanes, twirling lawn spinners, and moving the clouds around when they get boring. Harness the wind with today's Groupon to Live2kite in Greenbrae. Choose between the following options:

- For \$95, you get a two-hour introductory group kiteboarding lesson on land (a \$200 value).
- For \$165, you get a three-hour semiprivate kiteboarding lesson in water (a \$330 value). Students must have already taken a land-based lesson.

Seasoned wind harrassers enlighten students in the methods and maneuvers of kiteboarding through Live2kite's lineup of lessons for every skill level. Introductory classes of up to six beginners learn how to navigate the specialized board using a parachute-like sail. Students negotiate their provided crafts on land to prepare for future adventures in the ocean or outer space as experienced instructors break down the lingo, equipment basics, steering, and control of wind power. Those who have already mastered earthbound basics take to the Bay in semiprivate lessons with up to three other students. Gusty apprentices build upper-body strength with lessons in one-handed kite piloting, and more advanced boarders hone tricks and jumps developed during Benjamin Franklin's

**Live2kite**  
Company Website

---

**Groupon Now!**

- \$13 for a Large 1 Topping Pizza at Round...  
Redeem by 10pm
- \$5 for \$6 at Chast Cafe Express  
Redeem by 2pm
- \$5 for \$10 at Sunrise Deli  
Redeem by 2:30pm  
1 left

**View all Now! deals**

1,185 deals available today in San Francisco

---

**More Great Deals** See All

**Four-Hour Introduction to Glass Blowing 1, 2, or 3 Class at Revere Glass School in Berkeley Berkeley (Southwest Berkeley)**  
Over 630 bought

**\$99**  
\$200 value  
**View It!**

**San Francisco (Downtown)**  
\$195 for an In-Office Zoom! Teeth-Whitening Treatment at Sutter Dental (\$599 Value)  
Over 60 bought

**Multiple Locations**  
One-Way or Round-Trip SFO Airport Transportation from Bay Limousine (Up to 51% Off)  
Over 130 bought

**Mill Valley**  
\$25 for a Healthy-Chocolate-Making Workshop at Edible Goddess in Mill Valley (FREE Value)

# Our Relevance Scenario

Users



**GROUPON** Featured Deal | All Deals | Getaways | Goods | Reserve | Home-Garden | Gifts | Anywhere

Refer Friends. Get \$10\*

Search: Fitness, massages, etc. Location: San Francisco, CA

### AstaYoga – Mission Dolores

\$26 for Five Yoga Classes (\$70 Value)

from **\$25** Buy! **You Save 64%** ~~\$70~~ **\$45**

Give as a Gift | Learn more

Limited time remaining!

Over 770 bought  
Limited quantity available

The deal is on!

**In a Nutshell**  
The spacious, serene studio practices Ashtanga yoga and helps students achieve physical and spiritual benefits.

**The Fine Print**  
Expires 180 days after purchase. Limit 1 per person. May buy 1 additional as a gift. Limit 1 per visit. Must activate by reservation date. Reservation required. Classes must be used by same person. First class must be reserved in person. See the rules that apply to all deals.

**"Great Outdoors" Deals For You**

- \$14 for \$33 at Mount Granada Hotel and Lodge, Mount Shasta, CA
- \$104 for \$228 at Big Bear Lake National State Park, Big Bear Lake, CA
- \$103 for \$228 at Hotel Mountainaire, Tehama, CA
- \$10 for \$24 at Cascade Magician Subscription, Chico, CA
- \$24.99 for \$43 at Portland Outdoor Lights, Chico, CA

**GROUPON** Featured Deal | All Deals | Getaways | Goods | Reserve | Home-Garden | Gifts | Anywhere

Refer Friends. Get \$10\*

Search: Pets, massages, etc. Location: San Francisco, CA

### Citipets – Bayview

One or Three Days of Dog Daycare or One or Three Nights of Dog Boarding (Up to 51% Off)

from **\$20** Buy! **You Save 50%** ~~\$40~~ **\$20**

Give as a Gift | Learn more

Limited time remaining!

Over 50 bought  
Limited quantity available

The deal is on!

**In a Nutshell**  
Citipets and their staff-over-10 staff care for pets, who have access to communal sleeping rooms and play in a huge indoor-outdoor facility.

**The Fine Print**  
Expires 180 days after purchase. Limit 1 per person. May buy 1 additional as a gift. Limit 1 per visit. Valid only for certain purchases. Appointment required. 24-hour cancellation notice required. Must sign waiver. Not valid on holiday weekends. May require access to car. Dogs must meet all playground requirements listed here. See the rules that apply to all deals.

**"Great Outdoors" Deals For You**

- \$74 for \$124 at Mount Granada Hotel and Lodge, Mount Shasta, CA
- \$154 for \$208 at Big Bear Lake National State Park, Big Bear Lake, CA
- \$109 for \$228 at Hotel Mountainaire, Tehama, CA
- \$12 for \$24 at Cascade Magician Subscription, Chico, CA
- \$24.99 for \$43 at Portland Outdoor Lights, Chico, CA

**GROUPON** Featured Deal | All Deals | Getaways | Goods | Reserve | Home-Garden | Gifts | Anywhere

Refer Friends. Get \$10\*

Search: Fitness, massages, etc. Location: San Francisco, CA

### Regalito Rosticeria – Mission Dolores

Mexican Dinner for Two or Four with Appetizer, Entrees, and Dessert (Up to 47% Off)

from **\$30** Buy! **You Save 46%** ~~\$55.50~~ **\$25.50**

Give as a Gift | Learn more

Limited time remaining!

Over 800 bought  
Limited quantity available

The deal is on!

**In a Nutshell**  
Dishes in a super-hip, casual, farm-to-table kitchen. Also, more than 1000+ handmade pork, chicken, and seafood dishes. Cashless transactions in sales.

**The Fine Print**  
Expires 180 days after purchase. Limit 1 per person. May buy 1 additional as a gift. Limit 1 per table. Valid only for option purchased. Reservation required. Valid only Monday-Thursday. Must use appropriate vehicle to visit. See the rules that apply to all deals.

**"Great Outdoors" Deals For You**

- \$14 for \$33 at Mount Granada Hotel and Lodge, Mount Shasta, CA
- \$104 for \$228 at Big Bear Lake National State Park, Big Bear Lake, CA
- \$103 for \$228 at Hotel Mountainaire, Tehama, CA
- \$10 for \$24 at Cascade Magician Subscription, Chico, CA
- \$24.99 for \$43 at Portland Outdoor Lights, Chico, CA

**GROUPON** Featured Deal | All Deals | Getaways | Goods | Reserve | Home-Garden | Gifts | Anywhere

Refer Friends. Get \$10\*

Search: Pets, massages, etc. Location: San Francisco, CA

### K1 Speed – Multiple Locations

\$44 for a Racing Package with Four Races and Two Yearly Licenses (Up to \$91.96 Value)

from **\$44** Buy! **You Save 52%** ~~\$91.96~~ **\$47.96**

Give as a Gift | Learn more

Save left to shop 6 days 4:44-21

Over 610 bought  
Limited quantity available

The deal is on!

**In a Nutshell**  
Two-monthly events with access to up to 18 go-karts for solo and duos. Collection of racing memorabilia on display.

**The Fine Print**  
Expires Dec 4, 2013. Limit 2 per person. May buy multiple as gifts. Must activate license by registration date. License requires 1 year from activation date. All junior racers must be at least 10" and adult racers must be at least 5'11" tall to race. All items must be under 7" tall and 300g. See the rules that apply to all deals. Not valid for group events. May apply Groupon!

**"Great Outdoors" Deals For You**

- \$74 for \$124 at Mount Granada Hotel and Lodge, Mount Shasta, CA
- \$154 for \$208 at Big Bear Lake National State Park, Big Bear Lake, CA
- \$109 for \$228 at Hotel Mountainaire, Tehama, CA
- \$12 for \$24 at Cascade Magician Subscription, Chico, CA
- \$24.99 for \$43 at Portland Outdoor Lights, Chico, CA

# Deal Personalization Infrastructure Use Cases

**Deliver Relevant Experience with High Quality Deals**

**Deliver Personalized Website, Mobile and Email Experience**



**Deal Level Performance User Behavior Performance**



# Deal Personalization Infrastructure Use Cases

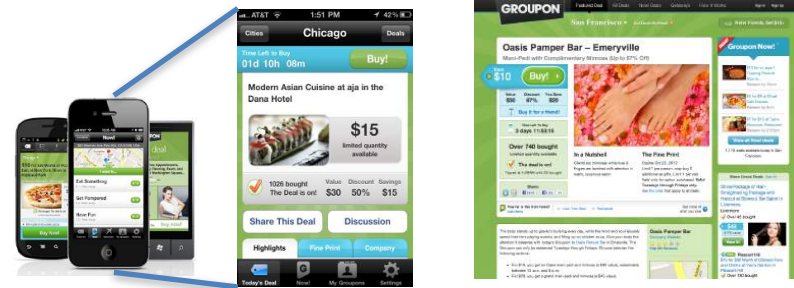
## Deliver Personalized Emails



Personalize billions of emails for hundreds of millions of users

### Offline System

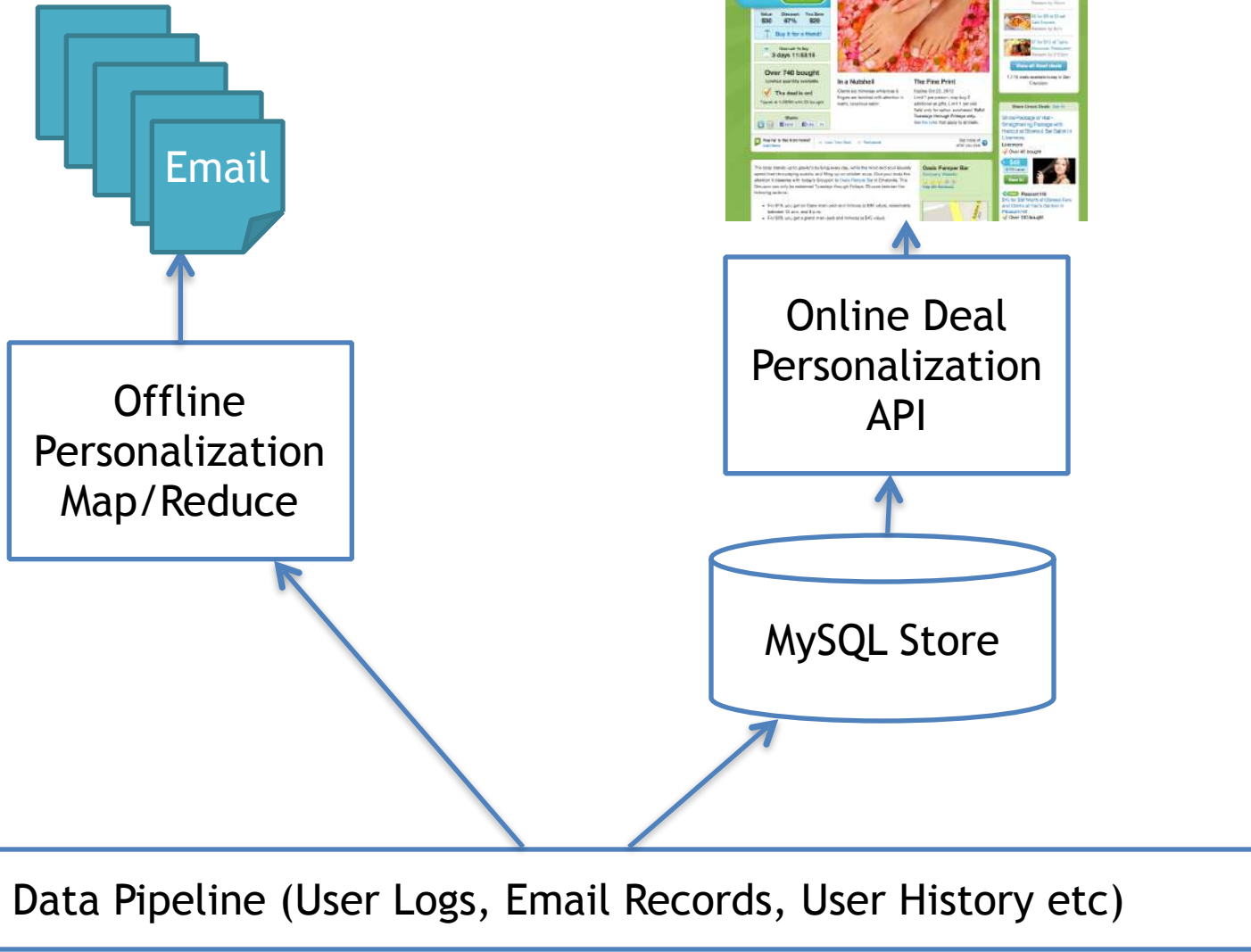
## Deliver Personalized Website & Mobile Experience



Personalize one of the most popular e-commerce mobile & web app for hundreds of millions of users & page views

### Online System

# Earlier System



# Scaling: Keeping Up With a Changing Business

## Growing Deals

2011



20+

2012



400+

2015



2000+

## Growing Users

- 110 Million+ subscribers
- We need to store data like, user click history, email records, service logs etc. This tunes to billions of data points and TB's of data



# Changing Business: Shift from Email to Mobile



110 Million+ App Downloads

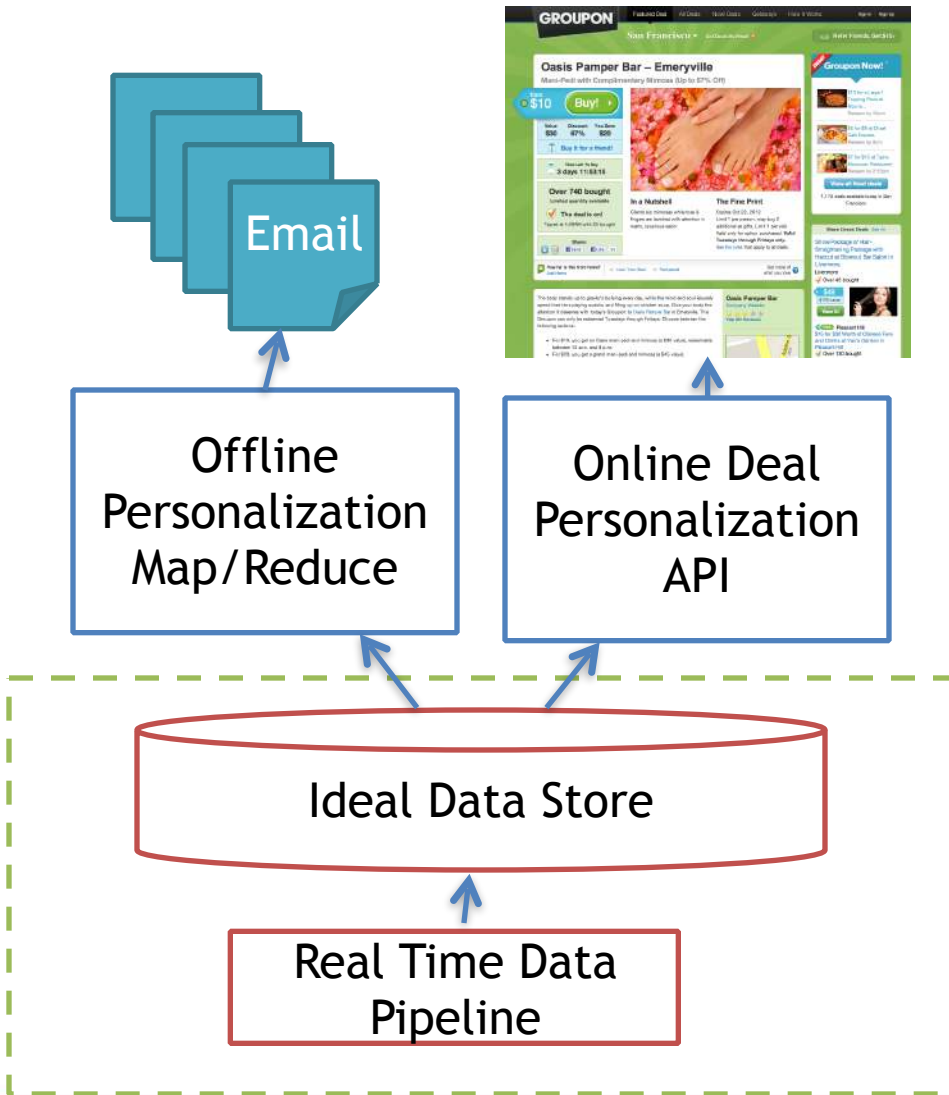
- Growth in Mobile Business
- Reducing dependence on email marketing
- Change in strategy from daily deal model to deal marketplace

# Earlier System



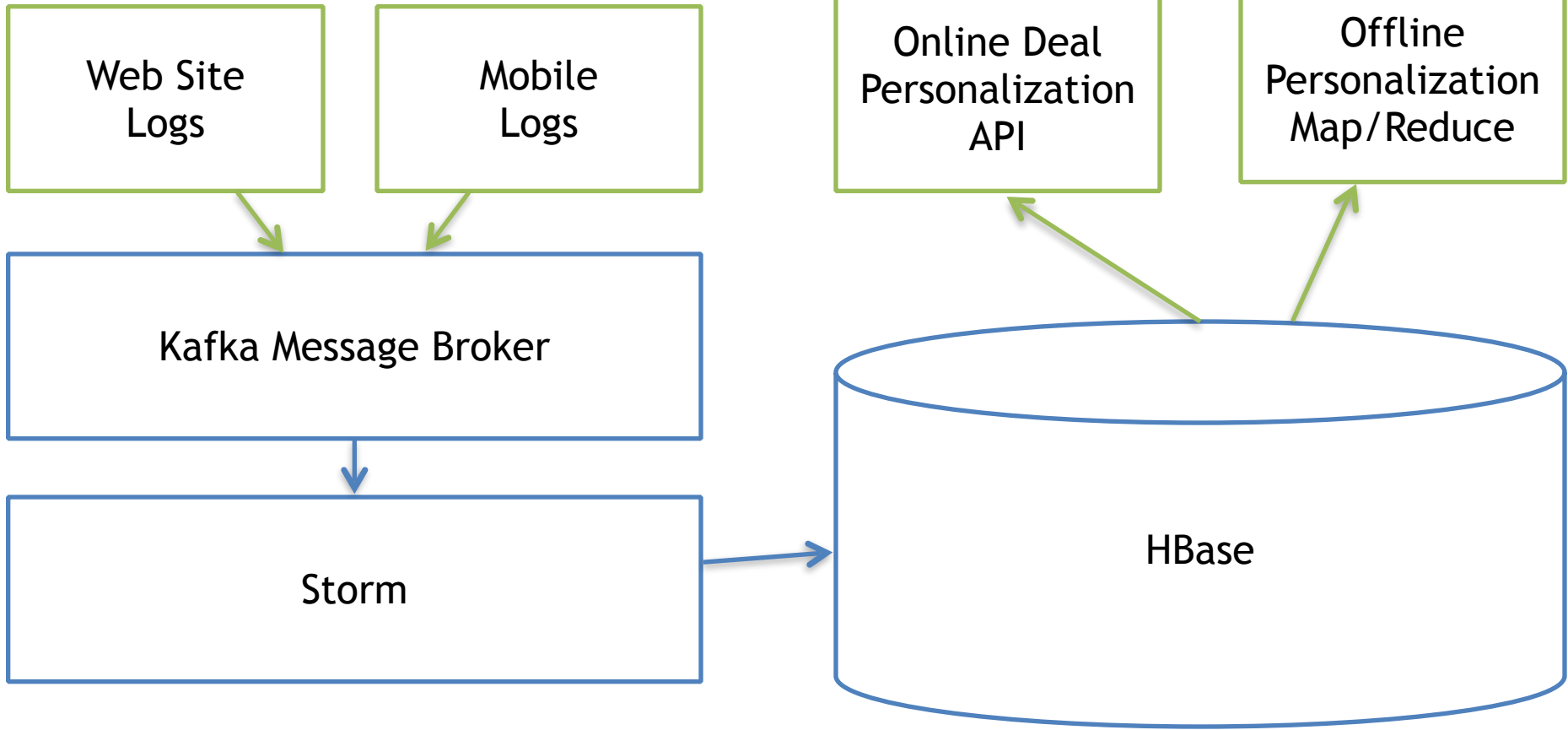
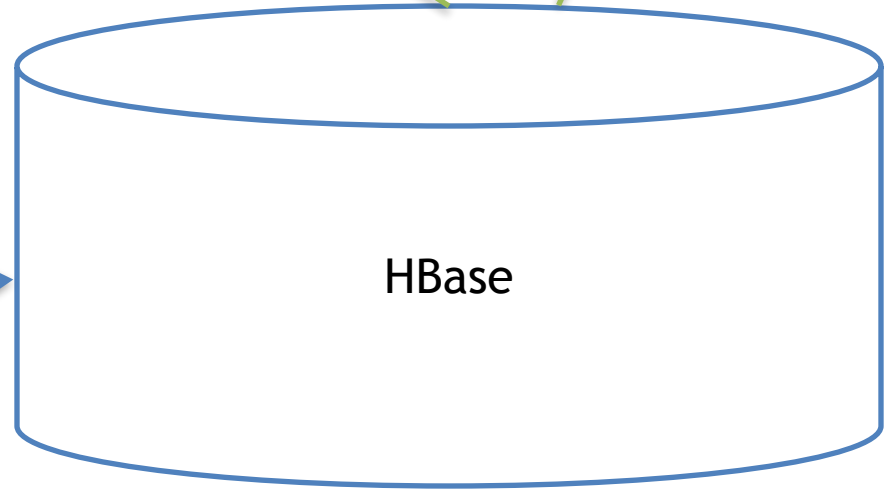
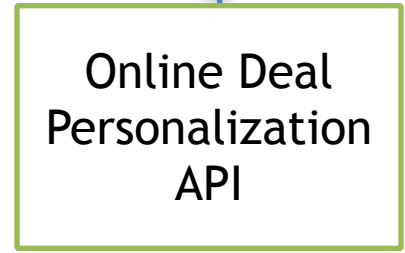
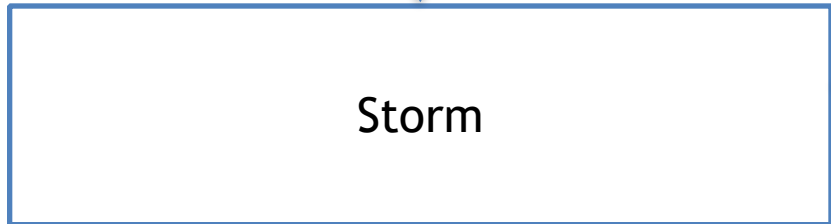
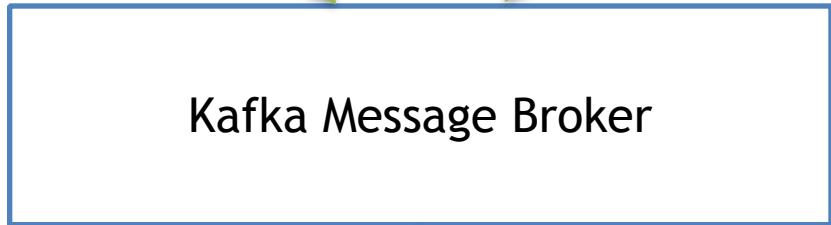
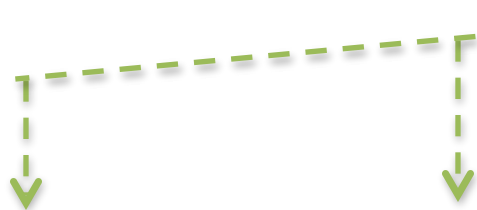
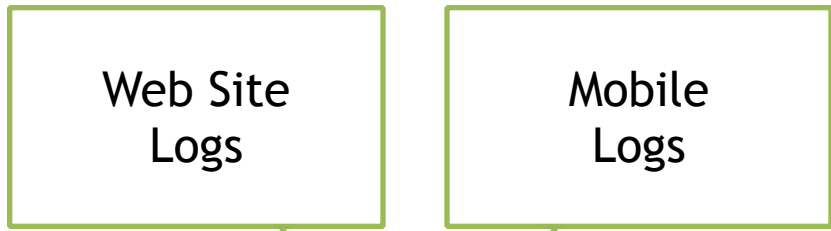
- Scaling MySQL for data such as user click history, email records was painful unless we shard data
- Data Pipeline is not “Real Time”

# Ideal System



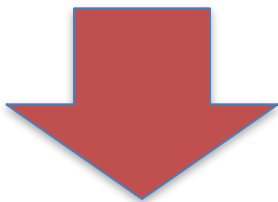
- Common data store that serves data to both online and offline systems
- Data store that scales to hundreds of millions of records
- Data store that plays well with our existing hadoop based systems
- Real Time pipeline that scales and can process about 100,000 messages/second

# Final Design



# Two Challenges With HBase

How to scale  
100,000  
writes/ second?



HBase

- How to run Map Reduce Programs over HBase without affecting read latency
- How to batch load data in HBase without affecting read latencies

HBase

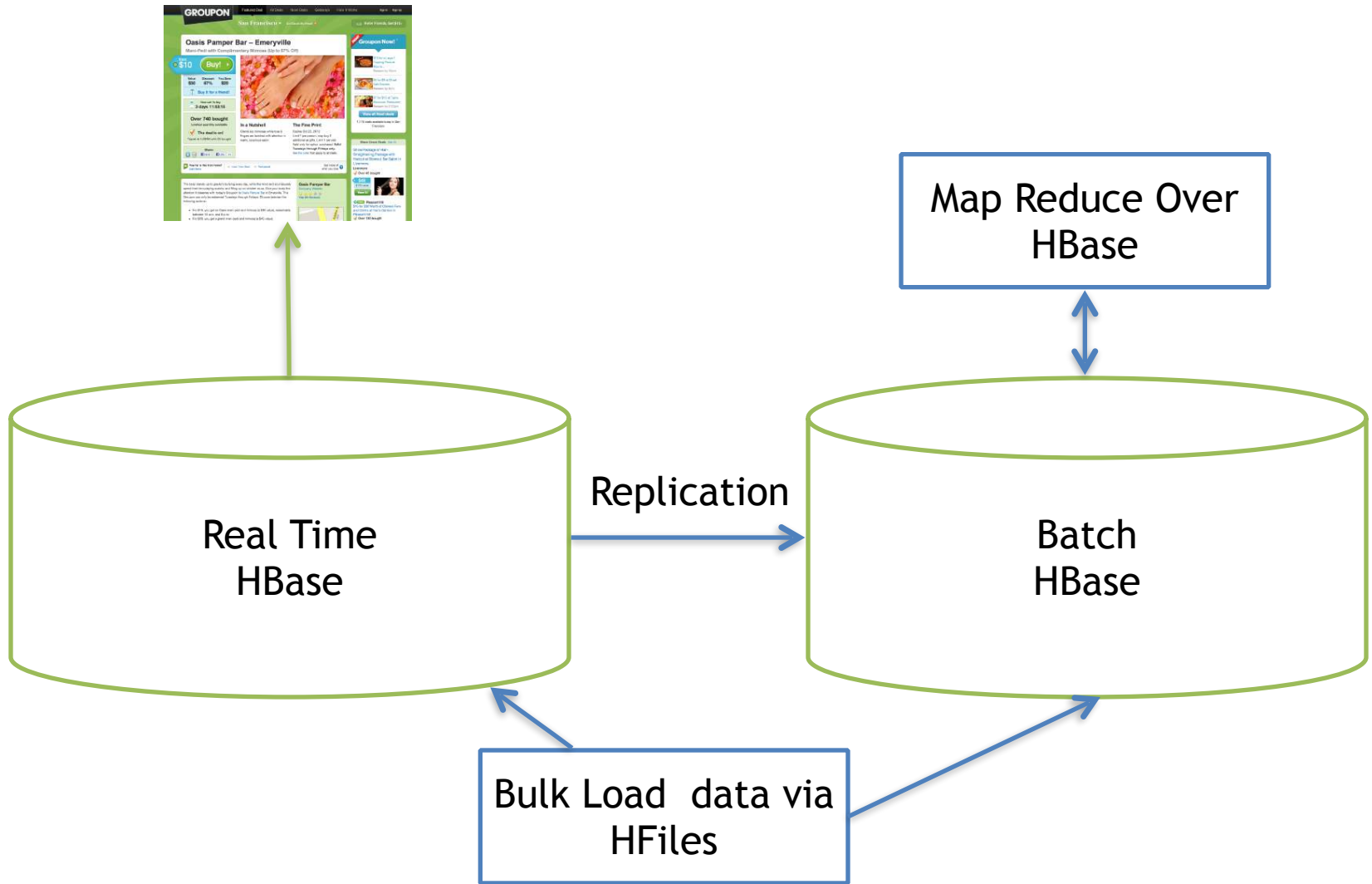
# Scaling with HBase

- Every new write is written as a new column in HBase. This also ensures de-duplication
- At the time of writing new event, no read happens

Column Family 0	
<i>User_id 1</i>	<i>timestamp_DealId_eventType: {Impression1}</i>
	timestamp_DealId_eventType: {Impression2}



# Final Design



# Leveraging System for Real Time Analytics

**Various requirements from relevance algorithms to pre-compute real time analytics for better targeting**



How do women in Berlin convert for Pizza deals?



How women in Berlin are converting for a particular pizza deal?

# Leveraging System for Real Time Analytics

## More Complex Examples



How do women in Berlin from Mitte area aged 45-50 convert for New York Style Pizza, when deal is located within 2 miles, and when deal is priced between €10-€20?



How do women in Berlin from Mitte area aged 45-50 convert for New York Style Pizza, for this particular deal which happens to be about 2 miles away, and deal is priced for €12

# Leveraging System for Real Time Analytics

## Even More Complex Examples

How do women in Berlin from Mitte area aged 45-50 convert for New York Style Pizza, when deal is located within 2 miles, and when deal is priced between €10-€20 who also like Activities such as Biking and who have been very active customer of Groupon deals on mobile platform?

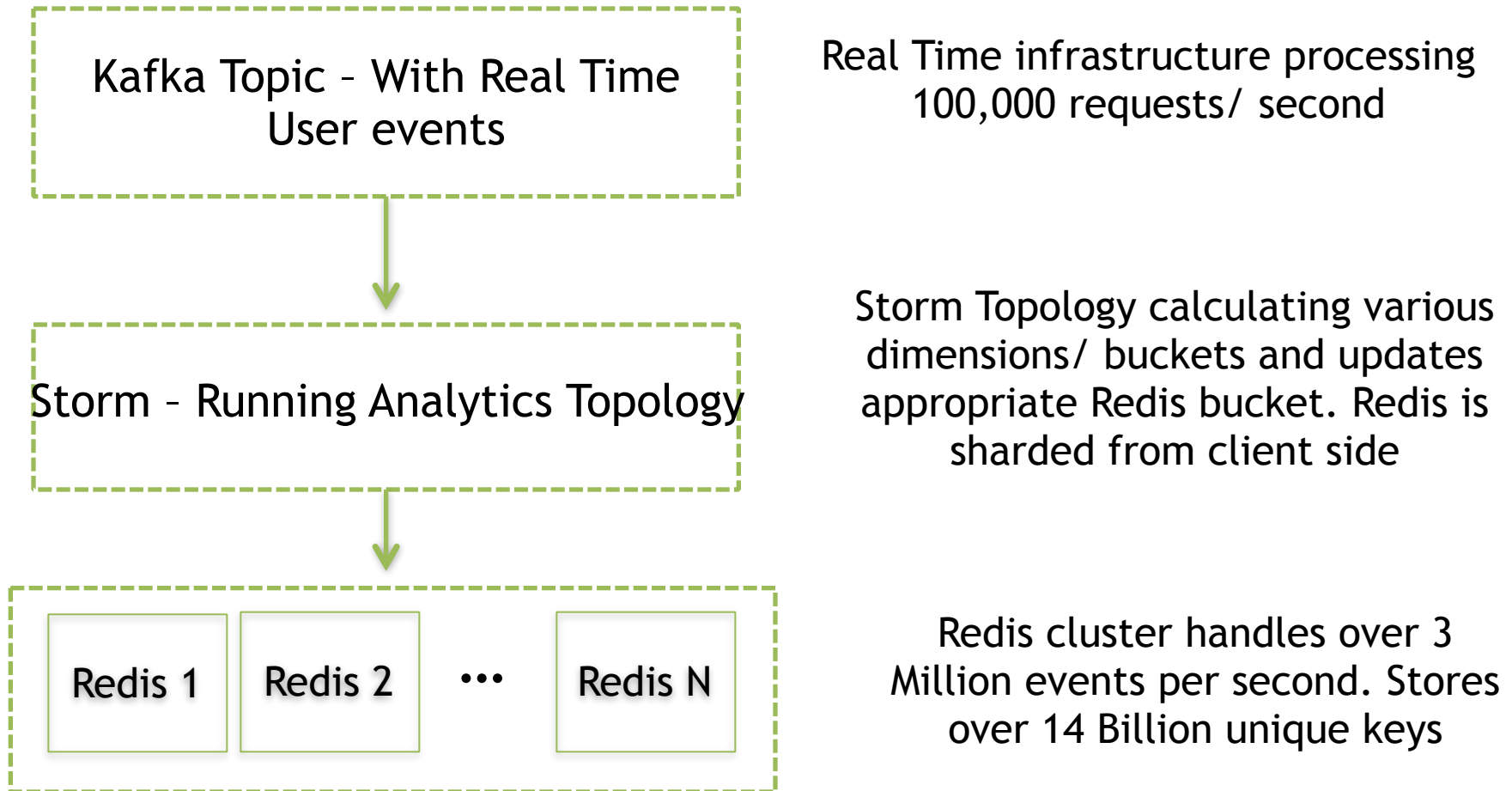
How do **women** in Berlin from **Mitte** area **aged** 45-50 convert for New York Style **Pizza**, for this **particular deal** which happens to be about **2 miles** away, and deal is **priced** for €12, who also like **activities** such as biking and who have been very **active customer** of Groupon deals on mobile **platform**?

# Power of Simple Counting

Turns out all earlier questions can be answered if we could count appropriate events in appropriate bucket

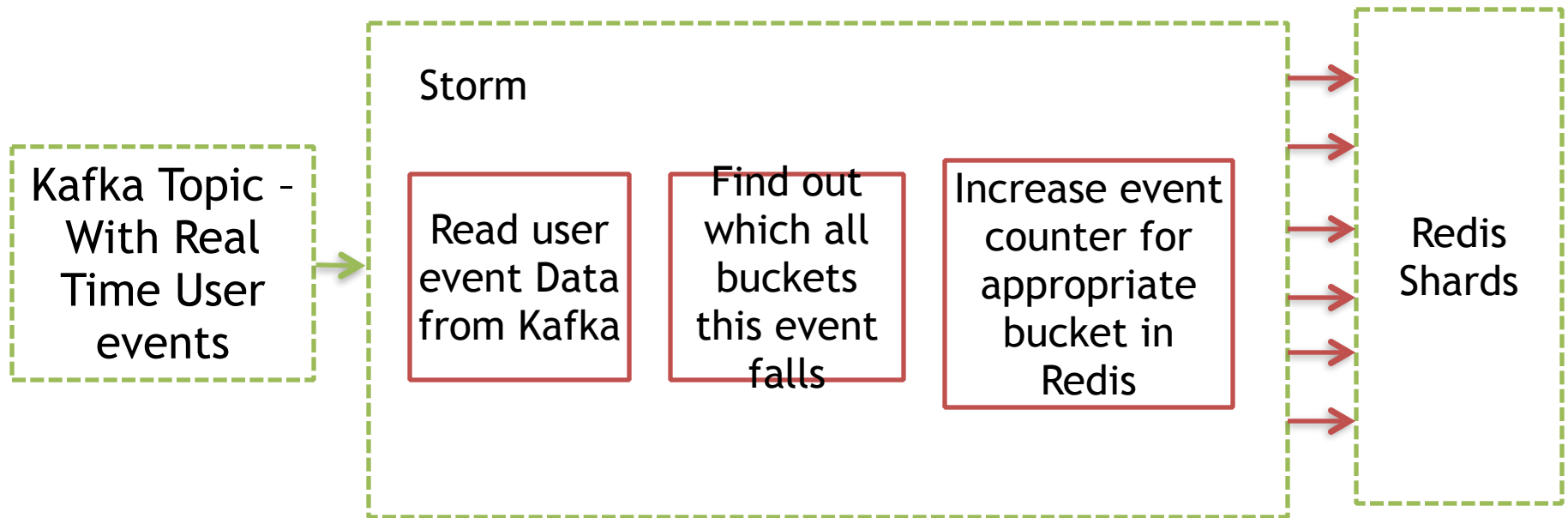
$$\text{Conversion rate for pizza deals for women in Palo Alto} = \frac{\text{No of Purchases by Women in Palo Alto for Pizza Deals}}{\text{No Deal Impressions by Women in Palo Alto for Pizza Deals}}$$

# Real Time Analytics Infrastructure





# Real Time Analytics Infrastructure - Explained



# Scaling Challenges - Kafka - Storm

- Massive reads and writes to Kafka cluster created back pressure in Storm topologies that produce data into Kafka
- Storm was hard to scale. We had to try various number of combinations to finalize how many bolts of each type are required for steady state operations and overall how many workers are needed etc. This took more time than we anticipated
- Use “maxSpoutPending” setting in Storm topologies. We found it to be very useful to shield your topologies from sudden increase in traffic
- Build your entire infrastructure - where data duplicates are allowed

# Scaling Challenges - Redis

- Reduce memory footprint - use hashes. Very memory efficient compared to normal Redis keys
- In order to support high write operations turned off AOF, turned on RDB backups

Easiest of all other infrastructure pieces - Kafka, Storm, HBase

# When Small is Big – Bloom Filters

- Since both Kafka and Storm can send same data twice specially at scale, it was important to build downstream infrastructure that can handle duplicate data.
- However, by very nature Analytics Topology (Counting Topology) cannot handle duplicates
- Storing individual messages for billions of messages is way too expensive and would take lot more memory
- So we used bloom filters. At a very small % error rate, we could effectively de dup data with a very small memory footprint. We store bloom filters in Redis using redis “bit” support

# Avoiding Errors – Backups/ Recovery Strategy

With such a high volume system, which also drives so much revenue for the company good backup/ recovery strategy is necessary

## Redis

RDB Backups every X hours. RDB backups are stored in HDFS for later use

## HBase

HBase Snapshot functionality is used. Snapshot taken every X hours. Can be loaded as necessary

## Kafka/ Storm

All input into Kafka topic is stored in HDFS. So any hour/ day can be replayed from HDFS if necessary

**We Are Hiring...**



# Questions?



Questions?

**GROUPON®**

Thanks!

[ameya@groupon.com](mailto:ameya@groupon.com)

[www.groupon.com/techjobs](http://www.groupon.com/techjobs)