data science and enterprise engineering

how data scientists and engineers work in tandem to achieve real-time personalization at overstock.com

overstock

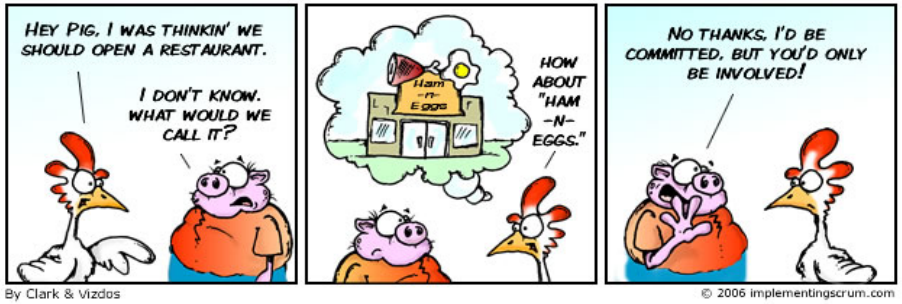
- pushing boundaries of retail since 1999
- Midvale, Utah
- 5 million unique products for sale
- billions of visits and page views
- 160+ countries

overstock marketing dev

remember animal farm?

team in the beginning.....

- pigs
  - data science team – 3 data scientists
  - engineering team - 2 developers, 1 QA, and a product owner
- chickens
  - plenty of business owners
  - lots of channel managers



overstock marketing dev

the problems we were up against

- data scientists were not working with the business to solve business problems
- engineers were not working with data scientists
- engineering was in a relational data mindset
- not regularly delivering business value
- solving yesterday's problems - today

days
minutes
seconds

business problem

This is where I add text. If I want it to keep going, it looks like this.

1st problem we came together to solve

- **real-time bidding** (RTB) is a means by which **advertising** inventory is bought and sold on a per-impression basis via real-time bidding, similar to financial markets
  - low-latency operation – need to bid >10ms
  - pre-compute scores nightly
  - push scored to cache in AWS
  - partnered with Bees Wax
  - replace existing 3rd party partners

problem to solve

This is where I add text. If I want it to keep going, it looks like this.

propensity to purchase

- identifying unique patterns and tendencies that indicate a user is ready to purchase
  - **user score**: represents a customer's likelihood of purchase
  - collected at the visit/user level
  - basic steps
    - turn raw user interactions into 'features'
    - train classifiers on months of data with label = purchase vs. no purchase
    - predict on new users/visits

challenges

data science

class imbalance

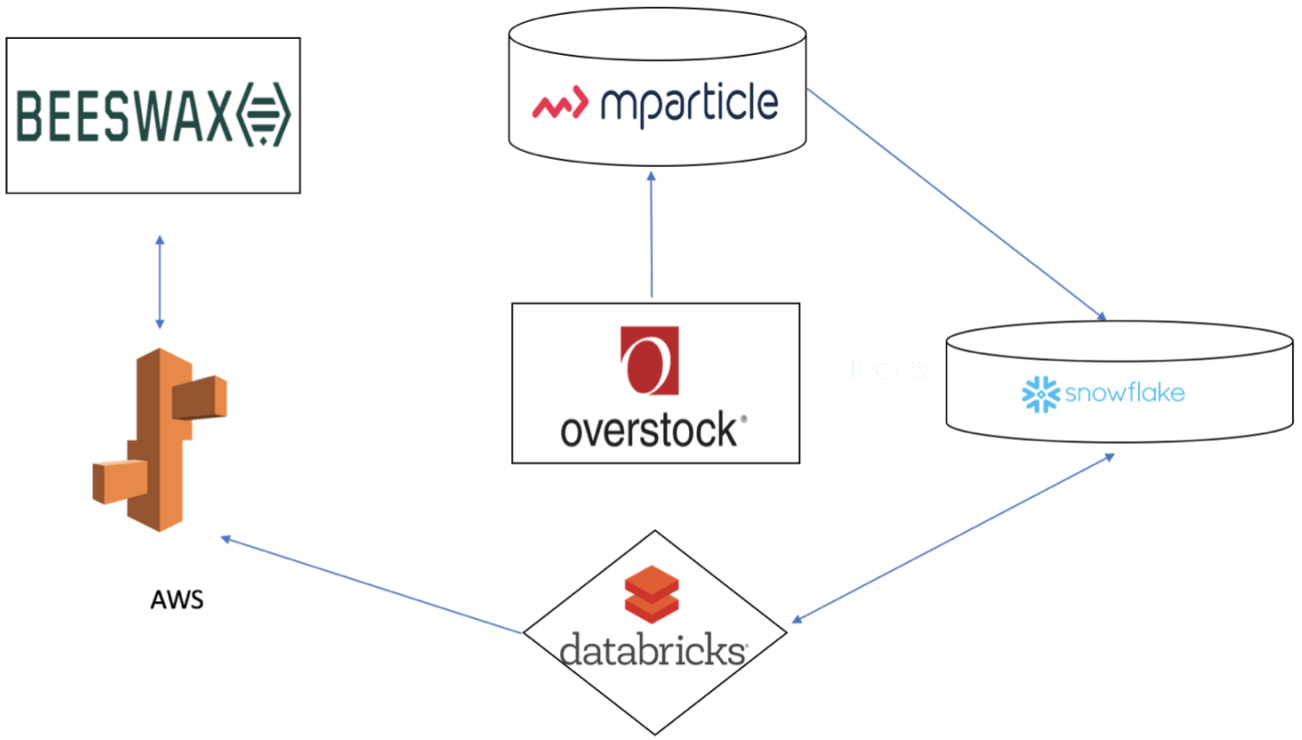
- **many new customers**
  - billions of unique page views in a calendar year
  - many users we are seeing for the first time (potentially)
- **low conversion**
  - a small percentage of sessions end in a purchase
  - sparse web logs mean we have to digest an enormous amount of data to generate useful features
- **we are interested in accuracy on the positive label**
  - recall over precision

challenges

what the team was up against

- **Constraints with current infrastructure and processes**
- **used to pulling data instead of streaming it**
- **data scientists are a scarce resource**
  - working on many non-relative tasks
  - not a lot of experience with the care and feeding of enterprise software
  - a bit set in the academia mindset

© Overstock



1

recap

Faster Scotty

what we accomplished in 6 months

- we were solving today's problems... by the end of the day
- engineering was thinking about streaming data
- had the pieces in place to start scoring users in minutes

days
minutes
seconds

business challenges

data science + engineering

needed to move faster

- score users faster
  - moved from daily to hourly
- picked up a fraud project
  - assign fraud score within 5 mins

challenges

data science + engineering

moving from daily batches to more regular micro-batches

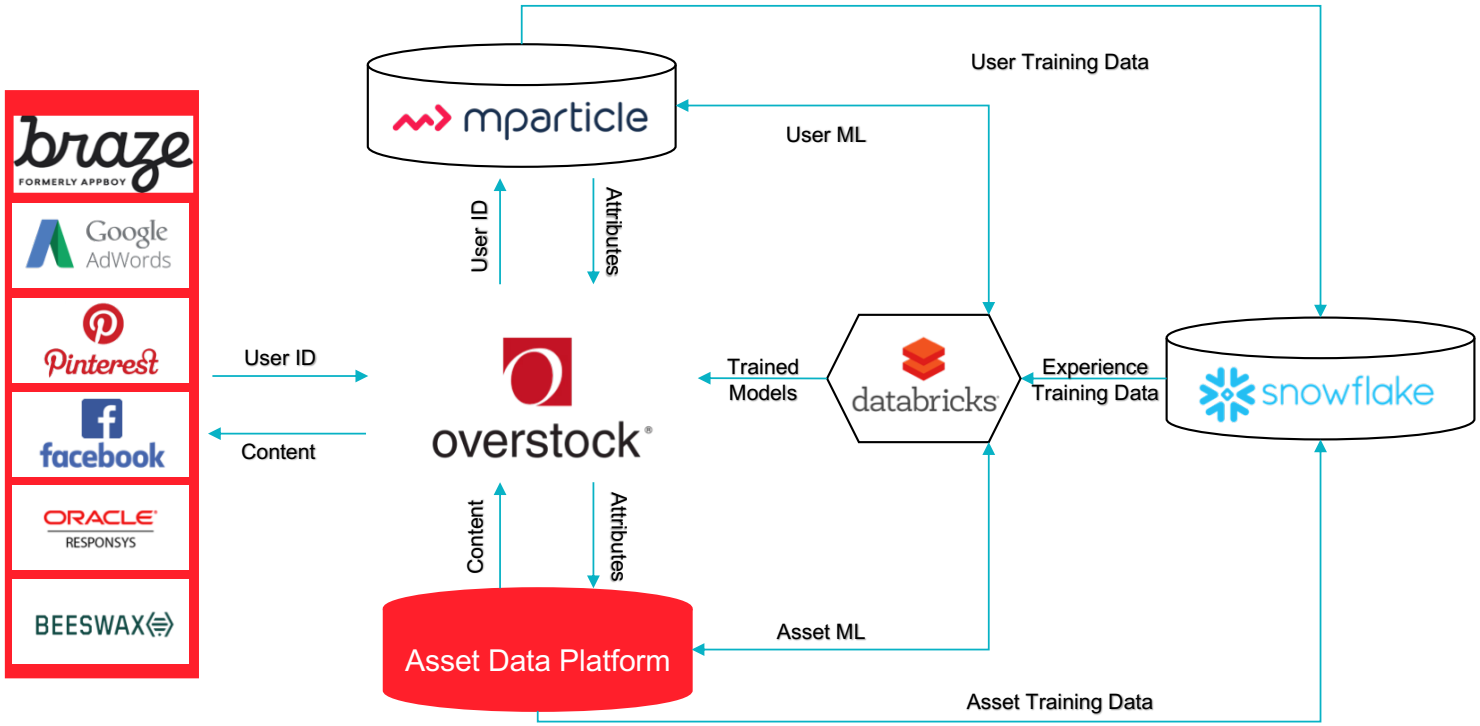
- **hourly jobs**
  - much more responsive scoring = larger server footprint
  - can still train offline in batch, but must have pipelines better tuned.
- **every 5 minutes**
  - ETL must be tightly honed
  - at this point you hit the edge of what is possible in the batch setting
  - feedback loops become critical for success

challenges

business vs engineering

balancing business goals with enterprise engineering

- as adoption increases visibility increases
  - need standardized data representations
  - have to make sure architecture is hardened and robust
  - need fail-safe mechanisms for critical processes
- MOST IMPORTANT
  - you can never, ever slow down overstock.com


braze
FORMERLY APPBOY
Google AdWords
Pinterest
facebook
ORACLE RESPONSYS
BEESWAX
User ID
Content
mparticle
User Training Data
User ML
User ID
Attributes
overstock
Trained Models
databricks
Experience Training Data
snowflake
Content
Attributes
Asset Data Platform
Asset ML
Asset Training Data

recap

data science + engineering

needed to move faster

- what was accomplished in 6 months
  - fraud scores in a minute or less
  - users were being scored in a minute
- what was left
  - still not fast enough for real time personalization on overstock.com

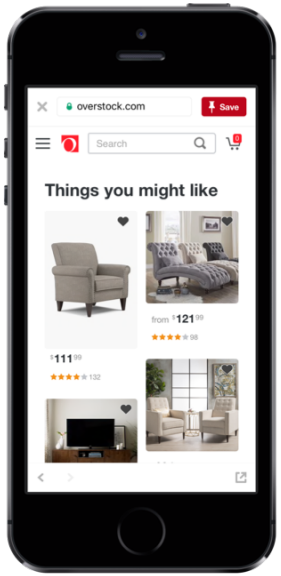
days
minutes
seconds

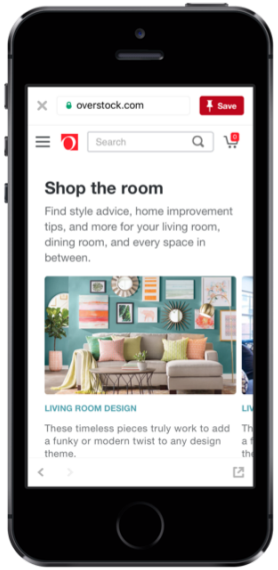
business problem

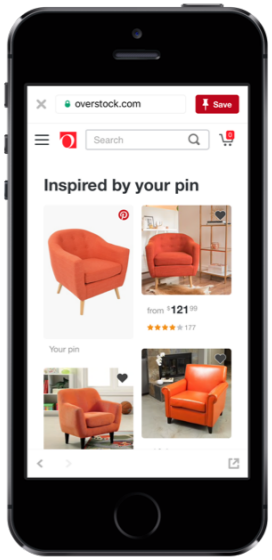
This is where I add text. If I want it to keep going, it looks like this.

near real-time personalization on the shopping site

- **near real-time personalization on overstock**
  - putting custom recs in front of the user.
  - can't slow the site down







overstock

challenges

data science +
engineering +
business

from micro-batches to near real-time

- near real-time can limit the size of models and complexity of calculations
  - we aren't trading stocks, we sell home goods
- we may not know much about a user we are trying to personalize for
  - what can you personalize for a new user as they initially interact with your site
    - heuristics moving towards full models

challenges

data science +
engineering +
business

from micro-batches to near real-time

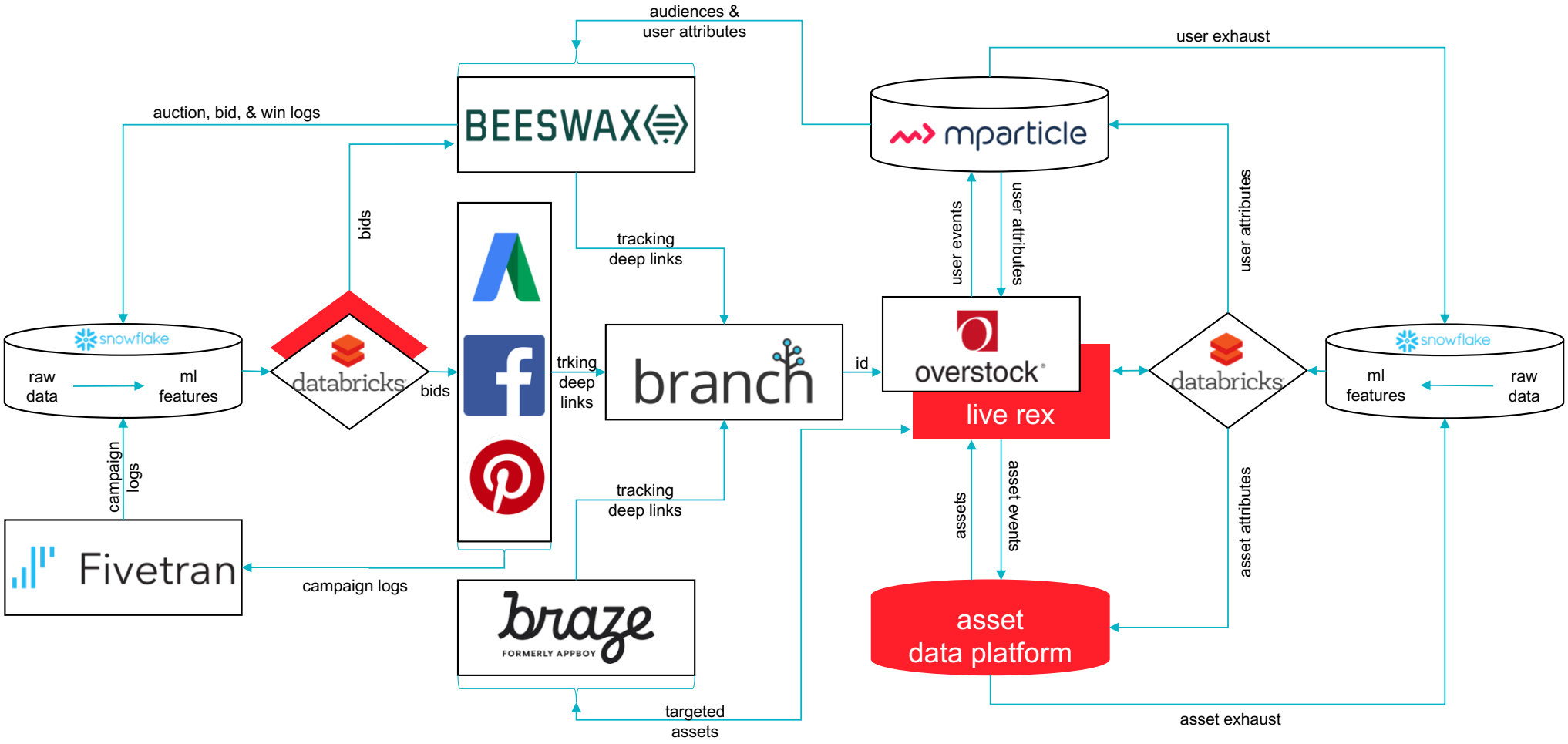
- real-time data flow requires rea-time analytics and strategy
  - introspection into processes becomes critical
  - must have automated process control to prevent algorithms from running wild
  - empower business owners to operate strategically without suffering from information overload

challenges

data science +
engineering +
business

from micro-batches to near real-time

- every piece of the process needs to function as a cohesive flow
  - do we need to wait for the data to get there, then act (fraud?)
  - or act on the data we have (recommendations for a shopping site)


audiences &
user attributes
user exhaust
auction, bid, & win logs
BEESWAX
mparticle
bids
tracking
deep links
user events
user attributes
user attributes
snowflake
raw
data
ml
features
databricks
bids
trking
deep
links
branch
id
overstock
live rex
databricks
snowflake
ml
features
raw
data
campaign
logs
tracking
deep links
assets
asset events
asset attributes
Fivetran
campaign logs
braze
FORMERLY APPBOY
asset
data platform
targeted
assets
asset exhaust

recap

data science +
engineering +
business

what was done and undone

- still not quite to near-real-time/low-latency recs for shopping site
- fraud score and email recommendations are fast enough

where we are going

data science +
engineering +
business

roadmap

- page-less personalization
- deep learning for fraud
- balancing real-time needs with efficient gains

We all need to be daring and collaborative to accelerate innovation!

take always

data science +
engineering +
business

lessons learned

- the cloud presents a whole new set of problems
- tech is hard, politics are harder
- we can’t operate in silos
- we must be aligned with the business
- the process must be iterative

questions?

and of course, we are hiring!

overstock™


overstock™