Conquering Big Data Complexity

JB Huang | VP, Data at GSN Games
Feb 10, 2016
Big Data Analytics Meet-up, Cambridge

Two Truths and a Lie

I once had so many Red Bulls I couldn't sleep for 3 days

My housemate plays guitar for Britney Spears

I've traveled to every continent
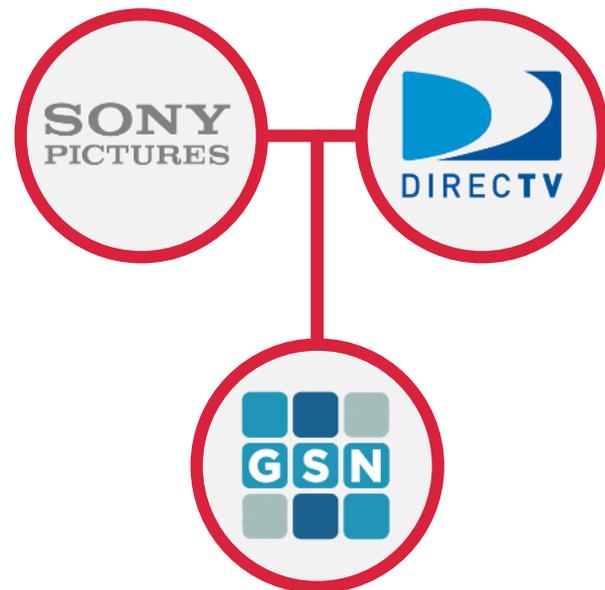
About Me

JB Huang

- VP, Data
- Joined in 2012
- Data & tech nerd
- Responsibilities: Data engineering, Reporting, Data visualization, Data science
- Prior industries: Education, Healthcare, Retail, Financial, Casino Gaming
- Love to travel – flying out to Dubai and India tomorrow
- From Boston area, currently "bi coastal"

GSN games

About GSN Games

Leading publisher of social casino games and cash tournaments

▶ Part of GSN, owned by Sony and DIRECTV

▶ Cable TV network in 80MM US households

▶ 100+ games on web, Facebook, iOS, Android

▶ 185MM global unique players

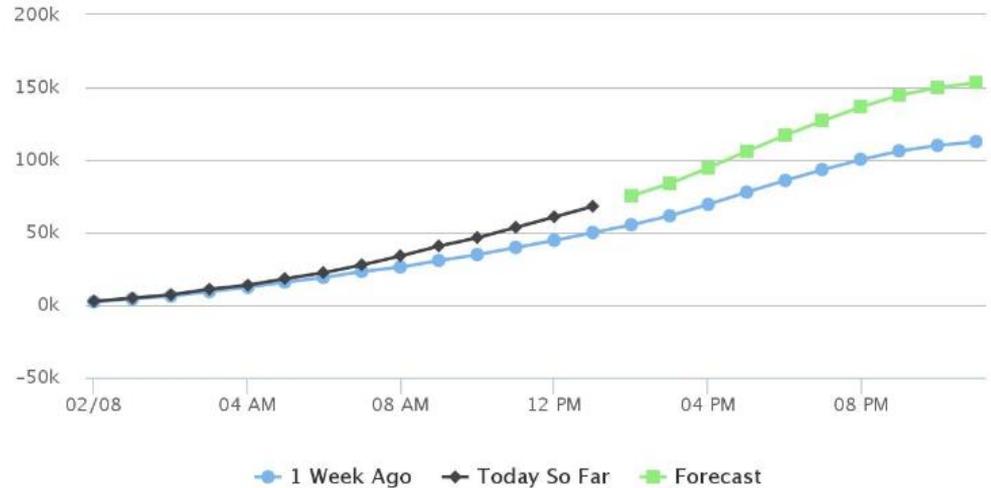▶ #6 and #10 top-grossing iPad apps of 2015

▶ #1 top-grossing solitaire app

How GSN Games accidentally became a "Big Data" company

Or, that one time we grew from 100,000 to 10,000,000 active users overnight

- **GSN is a very data-driven company**
  - 3+ billion events collected per day
  - 500TB data warehouse
  - 100% of employees actively use data
  - No silos: all data available to all staff
  - 50+ new A/B tests launched per week
  - Powers millions a year in user acquisition
  - All KPI tracking and forecasting
  - 100s of daily reports & dashboards

3 Complex Data Problems

Solved with tonight's sponsor →

Problem #1:

A/B Testing of New Mini Game

A/B test:

50% can play "Wheel of Fortune Bingo"

50% do not see game and can't play

Results:

Great Results on A/B Testing Dashboard!

Higher retention & revenue for app!

Conclusion:

New Mini Game is a Hit?

WebReports   KPIs   Games   Product   Tools   Search   Issues   jbhuang

A/B Test Dashboard

Important note: For boolean cohorts, you can just select the cohort. But for non-boolean cohorts, you must select which subcohort you want to use as the "enabled" group (to compare to "baseline").

Important note: In order to have new cohorts show up in the dropdown, please modify the Google Doc. After modifying the doc, it could take up to 3 hours for the new cohort to show up.

[1485] TripeaksPoiTreasureM... ▼   enabled ▼   Cohort Dates: 2016-01-08   2016-01-31   Activity Dates: 2016-01-08   2016-01-31   ☐ Advanced   **Run Report**

Metric	Baseline	Enabled	Difference	Confidence	Tools
All Devices					
ⓘ Ticketers	24	297	37.45%	86.90%	⏱ 00:45 SQL
ⓘ Devices	50,049	450,590	n/a	n/a	⏱ 00:32 SQL
ⓘ Paying Devices	1,565	14,855	n/a	n/a	⏱ 00:32 SQL
ⓘ Avg Bookings / Device	$ 0.42	$ 0.54	27.23%	100.0%	⏱ 00:35 SQL
ⓘ Avg Ad Revs / Device	$ 0.03	$ 0.03	5.15%	100.0%	⏱ 00:44 SQL
ⓘ Avg Days Active / Device	4.03	4.10	1.64%	99.84%	⏱ 00:41 SQL
ⓘ Device Pay %	3.13%	3.30%	5.43%	95.66%	⏱ 00:40 SQL
ⓘ Device Pay % (Tokens)	2.07%	2.16%	4.52%	82.93%	⏱ 00:40 SQL
ⓘ Device Pay % (Powerups)	2.12%	2.34%	10.40%	99.81%	⏱ 00:40 SQL
ⓘ Next Day Return Rate	42.81%	42.85%	0.10%	14.28%	⏱ 00:41 SQL
New Devices					
ⓘ Ticketers	19	243	41.60%	85.84%	⏱ 00:43 SQL
ⓘ Devices	35,723	322,659	n/a	n/a	⏱ 00:41 SQL
ⓘ Paying Devices	1,230	11,581	n/a	n/a	⏱ 00:32 SQL
ⓘ Avg Bookings / Device	$ 0.47	$ 0.58	22.24%	99.61%	⏱ 00:34 SQL
ⓘ Avg Ad Revs / Device	$ 0.03	$ 0.04	3.81%	98.72%	⏱ 00:39 SQL
ⓘ Avg Days Active / Device	4.42	4.48	1.31%	97.56%	⏱ 00:33 SQL
ⓘ Device Pay %	3.44%	3.59%	4.24%	84.15%	⏱ 00:32 SQL
ⓘ Device Pay % (Tokens)	2.29%	2.40%	4.93%	81.32%	⏱ 00:33 SQL
ⓘ Device Pay % (Powerups)	2.34%	2.54%	8.60%	97.86%	⏱ 00:27 SQL
ⓘ Next Day Return Rate	49.48%	49.30%	-0.34%	45.81%	⏱ 00:27 SQL

⬇ Download

GSN games

Problem #1:

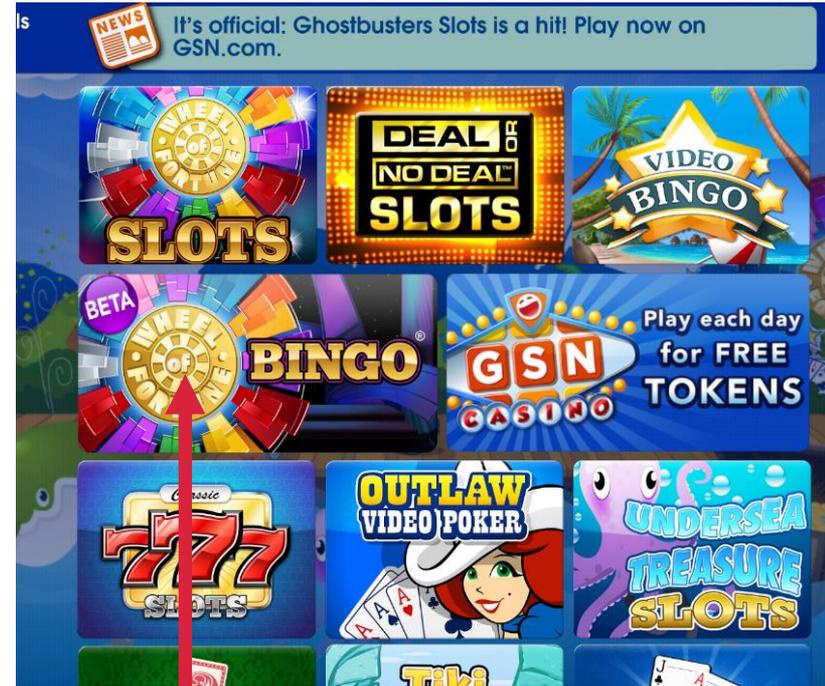**A/B Testing of New Mini Game**

Issue:

Very little revenue for actual mini game?

Actually:

Turns out logos were confusing and drove players to another game altogether.

Tools can simplify your results. Make sure you dig into the data as to why you got your results.
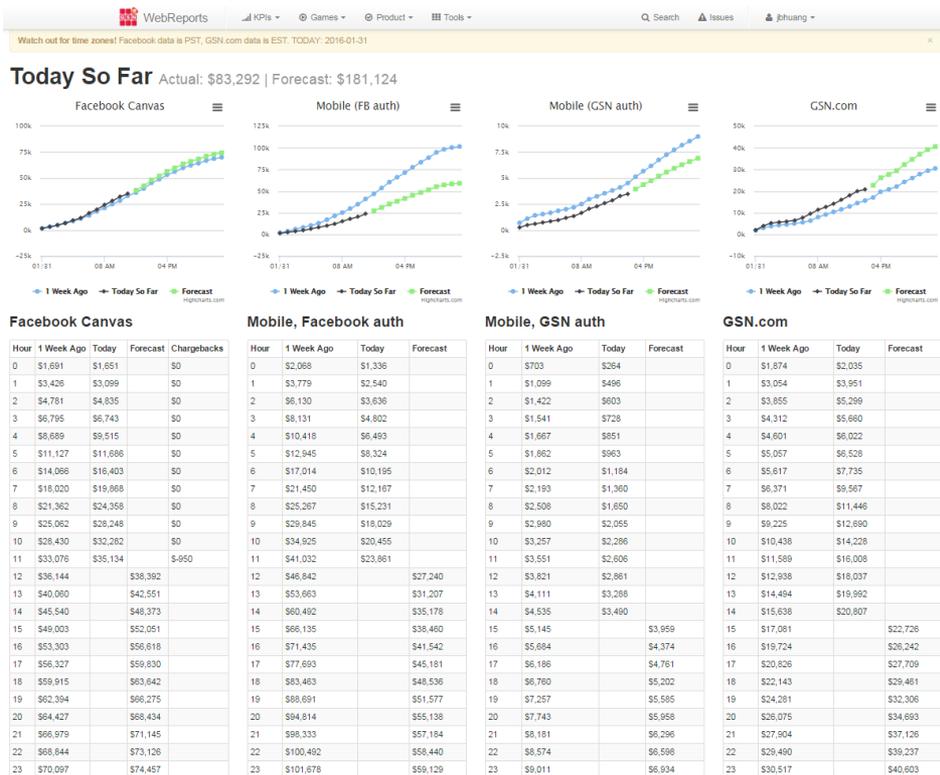
Problem #2:

Revenue Tracking in Mini Games

Track Revenue in Mini Games to pay our IP Partners.

Done using **free_amount & paid_amount**.
Fixed conversion of **$1 = 7500 paid_amount.**

What happens when someone breaks the fix conversion rate without realizing it?

We've over paid our IP Partners by millions over last 5 months!

Problem #2:

Revenue Tracking in Mini Games

*	event_time	fb_user_id	token_balance	paid_amount	free_amount	origin	game	paid_token_balance	amount_bet	amount_won
1	2015-12-13 00:01:02	100000478366981	20550	0	20000	returnbonus	NULL	0	(null)	(null)
2	2015-12-13 05:31:23	100000478366981	21600	0	1050	returnbonus	NULL	0	(null)	(null)
3	2015-12-13 05:32:30	100000478366981	40021600	5000000	35000000	payment	NULL	5000000	(null)	(null)
4	2015-12-13 05:34:15	100000478366981	39896100	-210000	84500	casino	Deal or No Deal Slots Mobile	4790000	210000	84500
5	2015-12-13 05:35:02	100000478366981	39496100	-400000	0	casino	Video Bingo Plus Mobile	4390000	400000	0
6	2015-12-13 05:40:14	100000478366981	39326100	-400000	230000	casino	Video Bingo Plus Mobile	3990000	400000	230000
7	2015-12-13 05:43:04	100000478366981	40591100	0	1265000	casino	Video Bingo Plus Mobile	3990000	0	1265000
8	2015-12-13 05:43:43	100000478366981	34466100	-3990000	-2135000	casino	American Buffalo Slots Mobile	0	6500000	375000
9	2015-12-13 05:45:04	100000478366981	33216100	0	-1250000	casino	American Buffalo Slots Mobile	0	1500000	250000
10	2015-12-13 05:45:44	100000478366981	28966100	0	-4250000	casino	American Buffalo Slots Mobile	0	4500000	250000

*	event_time	fb_user_id	token_balance	paid_amount	free_amount	origin	game	paid_token_balance	amount_bet	amount_won
1	2015-12-13 00:01:02	100000478366981	20550	0	20000	returnbonus	NULL	0	(null)	(null)
2	2015-12-13 05:31:23	100000478366981	21600	0	1050	returnbonus	NULL	0	(null)	(null)
3	2015-12-13 05:32:30	100000478366981	40021600	749925	39250075	payment	NULL	749925	(null)	(null)
4	2015-12-13 05:34:15	100000478366981	39896100	-210000	84500	casino	Deal or No Deal Slots Mobile	539925	210000	84500
5	2015-12-13 05:35:02	100000478366981	39496100	-400000	0	casino	Video Bingo Plus Mobile	139925	400000	0
6	2015-12-13 05:40:14	100000478366981	39326100	-139925	-30075	casino	Video Bingo Plus Mobile	0	400000	230000
7	2015-12-13 05:43:04	100000478366981	40591100	0	1265000	casino	Video Bingo Plus Mobile	0	0	1265000
8	2015-12-13 05:43:43	100000478366981	34466100	0	-6125000	casino	American Buffalo Slots Mobile	0	6500000	375000
9	2015-12-13 05:45:04	100000478366981	33216100	0	-1250000	casino	American Buffalo Slots Mobile	0	1500000	250000
10	2015-12-13 05:45:44	100000478366981	28966100	0	-4250000	casino	American Buffalo Slots Mobile	0	4500000	250000

Top is "dirty data" - Bottom is restated "clean data"

Problem #2:
Revenue Tracking in Mini Games

3 pass restatement of token adjustments table using payments received table.

200+ lines of SQL using moving window analytic sum & lag functions.

Lots of edge cases.

20+ hrs to write query
8+ hrs to restate 5 months

*	event_time	fb_user_id	amount_paid
1	2015-12-13 05:32:30	100000478366981	99.99
2	2015-12-13 10:31:57	100000478366981	19.99
3	2015-12-13 18:22:05	100000478366981	49.99

*	event_time	fb_user_id	token_balance	paid_amount	free_amount	origin
1	2015-12-13 00:01:02	100000478366981	20550	0	20000	returnbonus
2	2015-12-13 05:31:23	100000478366981	21600	0	1050	returnbonus
3	2015-12-13 05:32:30	100000478366981	40021600	5000000	35000000	payment
4	2015-12-13 05:34:15	100000478366981	39896100	-210000	84500	casino
5	2015-12-13 05:35:02	100000478366981	39496100	-400000	0	casino
6	2015-12-13 05:40:14	100000478366981	39326100	-400000	230000	casino

*	event_time	fb_user_id	token_balance	paid_amount	free_amount	origin
1	2015-12-13 00:01:02	100000478366981	20550	0	20000	returnbonus
2	2015-12-13 05:31:23	100000478366981	21600	0	1050	returnbonus
3	2015-12-13 05:32:30	100000478366981	40021600	749925	39250075	payment
4	2015-12-13 05:34:15	100000478366981	39896100	-210000	84500	casino
5	2015-12-13 05:35:02	100000478366981	39496100	-400000	0	casino
6	2015-12-13 05:40:14	100000478366981	39326100	-139925	-30075	casino

Problem #2:

Revenue Tracking in Mini Games

Data Integrity is still a complex data problem.

Data integrity checkers can help but you can never predict all the ways bad data can come in.

Sometimes you still have to get your hands dirty and do manual restatements and clean up.

Problem #3:

Tracking 1st User on Device

How do we figure out the first user associated w/ a mobile device?

We should have a "first user" event.

But business users never talked to engineers about tracking this.

app_name	synthetic_id	mesmo_id	created_at	source_event
GSN Casino	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	-275139723	2016-02-03 01:01:14	addDeviceMapping

*	event_time	event_name	attr29	mesmoid	synthetic_id	app_name
1	2016-02-03 01:01:11	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
2	2016-02-03 01:01:11	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
3	2016-02-03 01:01:12	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
4	2016-02-03 01:01:12	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
5	2016-02-03 01:01:12	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
6	2016-02-03 01:01:12	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
7	2016-02-03 01:01:13	addDeviceMapping	39A381DA-9B3B-4090-8DB6-9C6C0EE4A024	device	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	Mesmo Platform
8	2016-02-03 01:01:13	addDeviceMapping	F6C143F8-46B5-41F6-811D-2E80FC183D8B	device	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	Mesmo Platform
9	2016-02-03 01:01:13	addDeviceMapping	(null)	device	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	Mesmo Platform
10	2016-02-03 01:01:13	addDeviceMapping	B92CD369-64FA-4AA7-92F1-E9324697C583	device	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	Mesmo Platform
11	2016-02-03 01:01:13	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
12	2016-02-03 01:01:13	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
13	2016-02-03 01:01:13	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
14	2016-02-03 01:01:13	notifyError	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
15	2016-02-03 01:01:14	announcementEvent	(null)		013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino
16	2016-02-03 01:01:14	addDeviceMapping	-275139723	person	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	Mesmo Platform
17	2016-02-03 01:01:14	reportAnalyticEvent	(null)	(null)	013dfc1c-c12f-42f5-90d9-f1cacdfc0667	GSN Casino

GSN games

Problem #3:

Tracking 1st user on device

SQL queries to figure
this out took forever!

Vertica Pattern Matching
SQL to the rescue.

```sql
SELECT
    synthetic_id,
    event_time,
    event_name,
    attr29,
    app_name,
    event_name() as event,
    pattern_id() as pattern_id,
    match_id()
FROM gsnmobile.events
where event_time > sysdate - 3
MATCH
 (PARTITION BY synthetic_id ORDER BY event_time, event_name asc, attr30 asc
  DEFINE
    device AS event_name = 'addDeviceMapping' AND attr30 = 'person' AND synthetic_id is not null AND attr29 is not null,
    non1 AS event_name = 'addDeviceMapping' and attr30 <> 'person',
    non2   AS event_name <> 'addDeviceMapping' AND (app_name is null OR app_name = 'Mesmo Platform'),
    app    AS event_name <> 'addDeviceMapping' AND app_name is not null AND app_name <> 'Mesmo Platform'
  PATTERN
    P AS (device (non1|non2)* app)
  ROWS MATCH ALL EVENTS)
 ) a
WHERE event in ('device','app')
```

Having a conversation with business users to track the right data ahead of time can solves hours or days of data munging. But it's hard to predict every use case.

GSN games

Questions?

Contact Info:

✉

jb@gsngames.com

www.gsngames.com