Simply Business

Spark & GraphX Workshop

www.simplybusiness.co.uk

One Finsbury Square
London EC2A 1AE

Returning Visitors

The Challenge

Accurately identify users over time as they interact with us

Why? Knowledge

To understand how our visitors behave:

- They have long buying cycles
- They use multiple devices
- They engage with us through multiple channels

Why? Cost Analysis

To know how much it costs to acquire new customers:

- They visit us from different marketing channels
- Each of them have different costs
- We remunerate our partners depending on that

Why? Personalization

To adapt their visit while they interact with us:

- We can trigger offers depending on the channel they come from
- We can prefill fields if they’re recognized
- We can change the look and feel of the website

Cookies?

Their work great when you want to identify people coming back to your website.

Challenges:

- Different devices
- Cleared cookies
- Private sessions
- What about telephone calls…?

Use More IDs!

We can use other IDs provided by visitors to link sessions:

- Login details
- Email addresses
- Telephone numbers
- Credit card numbers
- Fingerprints
- …

Proof of Age Card

VALIDATE UK

Name/Enw
Your Name

DoB/Dyddiad Geni
04 May 1996

PASS

0101 1045 9843 2301

18+

ASSOCIATION OF CHIEF POLICE OFFICERS

ts trading standards institute

sia Security Industry Authority

Example

How many visitors do we have here?

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Example

How many visitors do we have here? Only 2

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111, a@a.com

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111, a@a.com
Visitor B: 555

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111, a@a.com
Visitor B: 555
Visitor C: 888

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111, 555, a@a.com
Visitor C: 888

Visitors A and B were merged!

Example: Streaming

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Visitor A: 111, 555, a@a.com
Visitor C: 888, b@b.com

Batch Solution

Example: Batch

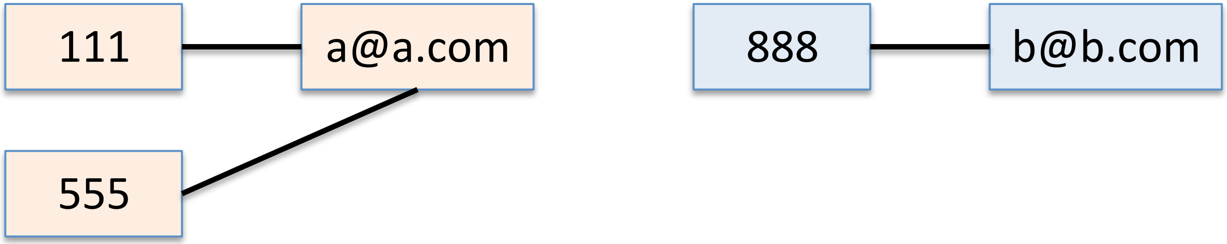
Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com

Batch Algorithm

- Find the connected components in the graph
- Already implemented in GraphX!
  - Vertices: visitor IDs
  - Edges: pairs of IDs occurring in the same event

Example: Batch

Timestamp	Event	Cookie ID	Email
1	page_view	111	
2	details_submitted	111	a@a.com
3	page_view	555	
4	page_view	888	
5	policy_sold	555	a@a.com
6	details_submitted	888	b@b.com



GraphX: Issues

- There is no Python API for GraphX, and Graphframes performance is worse than GraphX's
- GraphX requires vertices to have numeric IDs
    - You'll have to create a mapping between the real IDs and the numeric IDs
    - Perform the connected components calculation
    - Map back the numeric IDs to the real IDs
- Documentation is not as good as other Spark projects

Beware

That two events share an ID does not necessarily mean that were generated by the same person. Watch out for:

- Clashing IDs
- Fake IDs: test@test.com, 07123456789, etc.
- Shared devices: couples, public computers, etc.
- People using your system on behalf of someone else
- Bugs

Analyze your data first!

Whizz-Kidz

move a life forward

Whizz-Kidz & Simply Business

Whizz-Kidz provides disabled children with the essential wheelchairs and other mobility equipment they need to lead fun and active childhoods.

Simply Business and its employees have pledged to raise £150,000 during the next three years. This amount of money is enough to clear Whizz Kidz' waiting list in both the London and Northampton region!

https://simplybusiness.everydayhero.com/uk/sahara-2016

Thank You!