
MMXIV
HACKATHON

WE ARE BACK

on the line

- wifiSSID: hackathonclt
- Password: no password (like really there is no password)

@charlottehacks

#HACKATHONCLT

agenda

friday, april 25

5p	Registration Table Opens
6- 8p	Kickoff & Party
8p	Hack Problem Presentation
8:30p	**Go Hack**
12a	**Midnight Snacks**

saturday, april 26

7a-8a	Breakfast
12p	Hackathon Ends	Lunch
1:30p	Presentations & Judging Begin
2:30p	Awards Ceremony

the basics

- what is a hackathon?
- why charlotte?
- what's the motive?
- big data AGAIN?

what made it possible

- community
- community
- community

who made it possible

rules of engagement

- nothing illegal
- respect copyright
- keep it clean
- terms & conditions
- all work must be on site
- organizers reserve the right…

THE PROBLEM

business problem

using actual shopper behavior and product categories, how can a leading retailer, that is one of P&G's premier partners, better predict Online Shopping proclivity at the individual customer level

what are you looking for

- understanding of the data (it is big)
- product usage
- current online channel usage
- ‘center-store’ product linkage
- triggers that indicate shift from offline to online
- scale trumps complexity in algo

how will you share results

- format - go wild…as long it is not a powerpoint
- time – will depend on number of entries…we will divide the 90 minutes equally amongst
- creativity, predictability, scalability are all important
- judges will be revealed tomorrow… they are hard but they are awesome

THE DATA

what you get

- real-world retail data
- 558,574,302 records
- scrubbed & de-identified
- 18 field names
  - Customer ID (masked)
  - transaction amount
  - Item code
  - Item quantity
  - Discounts (if any)
  - Online shopping code

Field Name	# Unique Values	# Null Values	# Records	Example	Description
HHID	642,141	-	558,574,302	7a174beb-ab24-4181-97b7-20c9c9bdac47	Unique Household ID
ITEM_DESCRIPTION	76,424	-	558,574,302	BANANAS, YELLOW	Text description of the item purchased
UPC_NUMBER	96,994	-	558,574,302	4011	Unique item ID
MASTER_UPC_NUMBER	52,552	200,791,320	558,574,302	7203663995	Unique master item ID
SUBCATEGORY_DESCRIPTION	2,285	-	558,574,302	DIET	Text description of the subcategory purchased
SUBCATEGORY_NUMBER	2,232	-	558,574,302	54	Unique subcategory ID
CATEGORY_DESCRIPTION	368	-	558,574,302	FRESH PRODUCE	Text description of the category purchased
CATEGORY_NUMBER	335	-	558,574,302	64	Unique category ID
DEPARTMENT_DESCRIPTION	26	-	558,574,302	G1 GROCERY	Text description of the department purchased
DEPARTMENT_NUMBER	22	-	558,574,302	1	Unique department ID
RECEIPT_NUMBER	32,412,750	-	558,574,302	1373861190	Unique visit ID
ITEM_QUANTITY	328	-	558,574,302	1	Number of itema purchased
DISCOUNT_QUANTITY	32	-	558,574,302	1	Number of items purchased on discount
EXTENDED_PRICE_AMOUNT	29,816	-	558,574,302	3.99	Price of item purchased - pre discount
EXTENDED_DISCOUNT_AMOUNT	9,362	-	558,574,302	1	Discount of item purchased
TENDER_AMOUNT	58,290	-	558,574,302	1	Amount paid per receipt number (including cash backs and/or change due)
TRANSACTION_DATETIME	290,064	-	558,574,302	11/19/13 15:40	Time of the transaction
EXPRESS_LANE	2	-	558,574,302	1	Flag identifying whether the transaction is (1) or is not (0) an express lane transaction

TECH

hardware stack

- At scale hadoop cluster (hat tip DataChambers)
- 100 cores
- 800GB RAM
- 20TB storage

dataset stored

- stored in HDFS
- data dictionary available on github here: http://www.github.com/tresata/hackathonclt
- DO NOT PULL DOWN THE ENTIRE SET
- DO NOT LEAVE WITH ANY DATA
- YES WE WILL SPOT CHECK

languages

- anything that can compile to a .jar (like java, scala, etc)
- JDBC connection available through hive
- Python, via pyspark

wifi

- SSID: hackathonclt
- Password: (you know it)

support

- @charlottehacks | #hackthonCLT
- look for black ‘staff’ shirts

THE PRIZE

in it to win it

$4,000 GRAND PRIZE

$1,000 RUNNER UP

GOOGLE GLASS JURY PRIZE

LOGISTICS

agenda

friday, april 25

5p	Registration Table Opens
6- 8p	Kickoff & Party
8p	Hack Problem Presentation
8:30p	**Go Hack**
12a	**Midnight Snacks**

saturday, april 26

7a-8a	Breakfast
12p	Hackathon Ends	Lunch
1:30p	Presentations & Judging Begin
2:30p	Awards Ceremony

the basics

- restrooms – located on 5th and 4th floors
- building access – DO NOT LEAVE
- be reasonable – alcohol, wristbands, etc.
- lounges – vote for your favorite tomorrow
- support – look for black shirts

questions?

- Katie Levans – Event
- Aussie Jack – Tech
- Richard Morris – Data
- Chase & Abhi – Business Problem

format

- teams no larger than 2
- pick a team name
- ideal composition – coder, designer, presenter
- hacking ends at 12p Saturday 4/26
- presentations begin at 1:30p

LET'S HACK

@charlottehacks | #HACKATHONCLT

http://www.github.com/tresata/hackathonclt