



R CASE STUDY FROM EBAY APD

李忠

zholi@ebay.com

AGENDA

- EBAY APD Introduction
- R@EBAY APD
- Site Speed Case Study
- Buyer Segmentation Case Study
- Lesson and Learn
- Q & A

- All Categories >
- Fashion >
- Motors >
- Electronics >
- Collectibles & Art >
- Home, Outdoors & Decor >
- Entertainment >
- Deals & Gifts >
- Sporting Goods >
- Classifieds



Welcome to eBay

Whether you're new to eBay or a veteran user, we have just the right tools to get you on the right track.

-  [New to eBay](#)
-  [How to buy](#)
-  [How to sell](#)
-  [Increase your sales](#)

Shop safely on eBay



eBay Buyer Protection
We've got you covered!



eBay Top-Rated Sellers
Get great service & fast shipping from top-rated sellers.



PayPal
PayPal is the world's most-loved way to pay and get paid.

Sign in

Back for more fun? Sign in now to buy, bid and sell, or to manage your account.

[Sign in](#)

Not registered yet?

Join the millions of people who are already a part of the eBay family.

[Register](#)

THANK YOU
FOR SHOPPING AT


Tech favorites and best sellers



Apple iPod touch



Canon Digital Cameras



Nintendo Wii



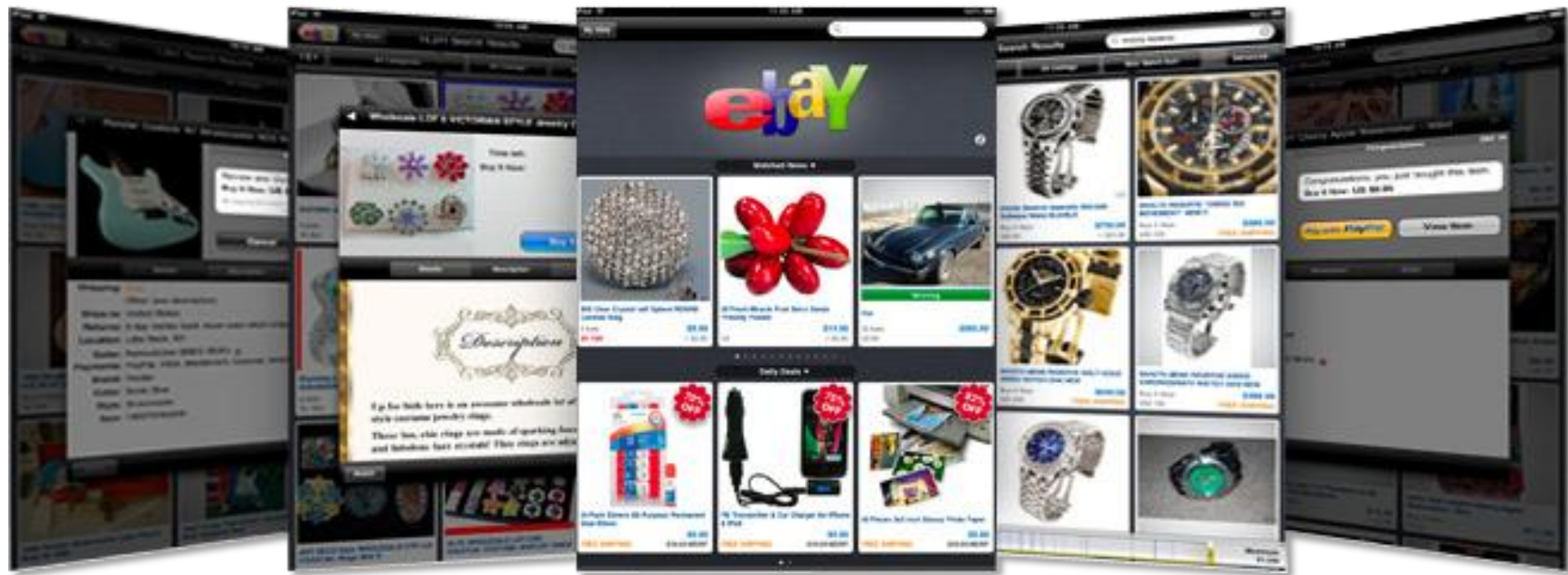
HTC EVO 3D

eBay stories  

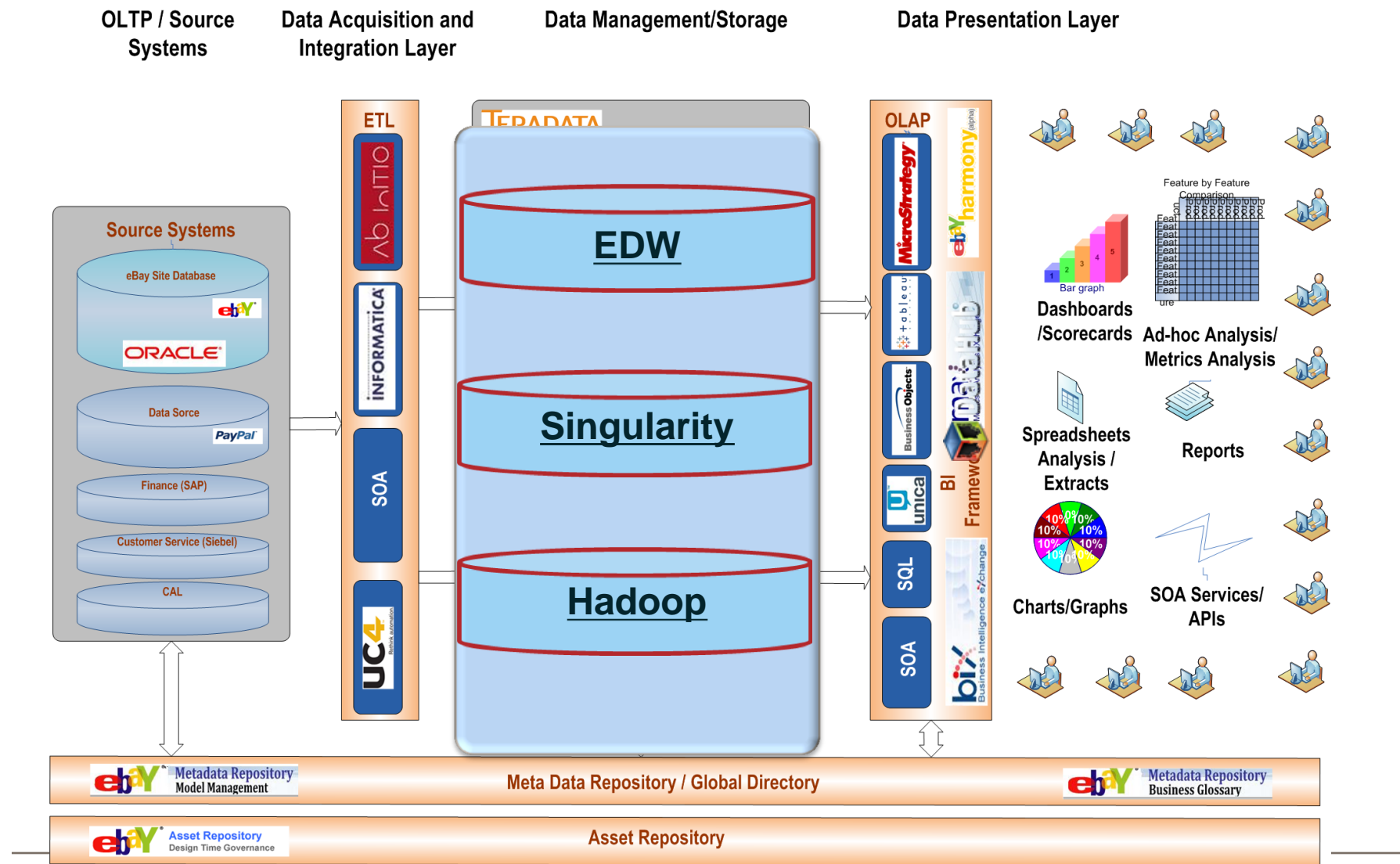
Bon Iver's whiskey barrel guitar

Updated Tue, Oct 23, 2012

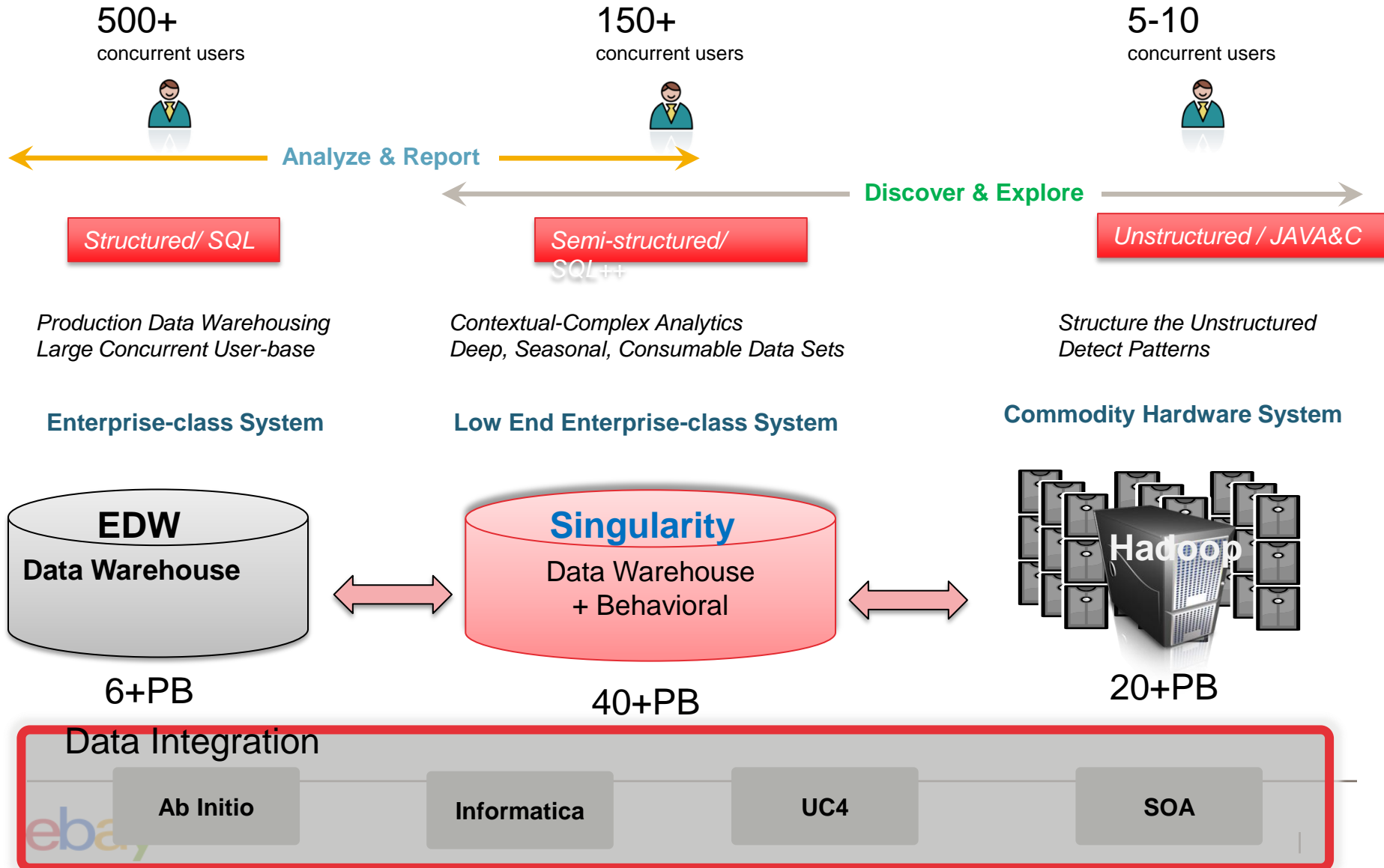
In this day of mass-produced guitars, it is nice to know there is still a niche business catering to individual guitar-making. But a guitar made from whiskey barrels? You need to see it to believe it. [Continue reading](#) —



Analytics Platform Architecture



Data Platforms




R@EBAY APD


- 2010 – Adopt R
- 2011 – Kick off R Cop
- 2012 – Integrate R into Insight Product
- 2012 - Kick off Data Science Cop





CoP For R created on Thursday, 12 January 2012

 Edit

 Like 0

 Bookmark

 Leave Group

 Invite Member

Community of Practice for R

Recent Announcements >



The China Fifth R Conference Meeting Registration is starting now... (1 NEW)

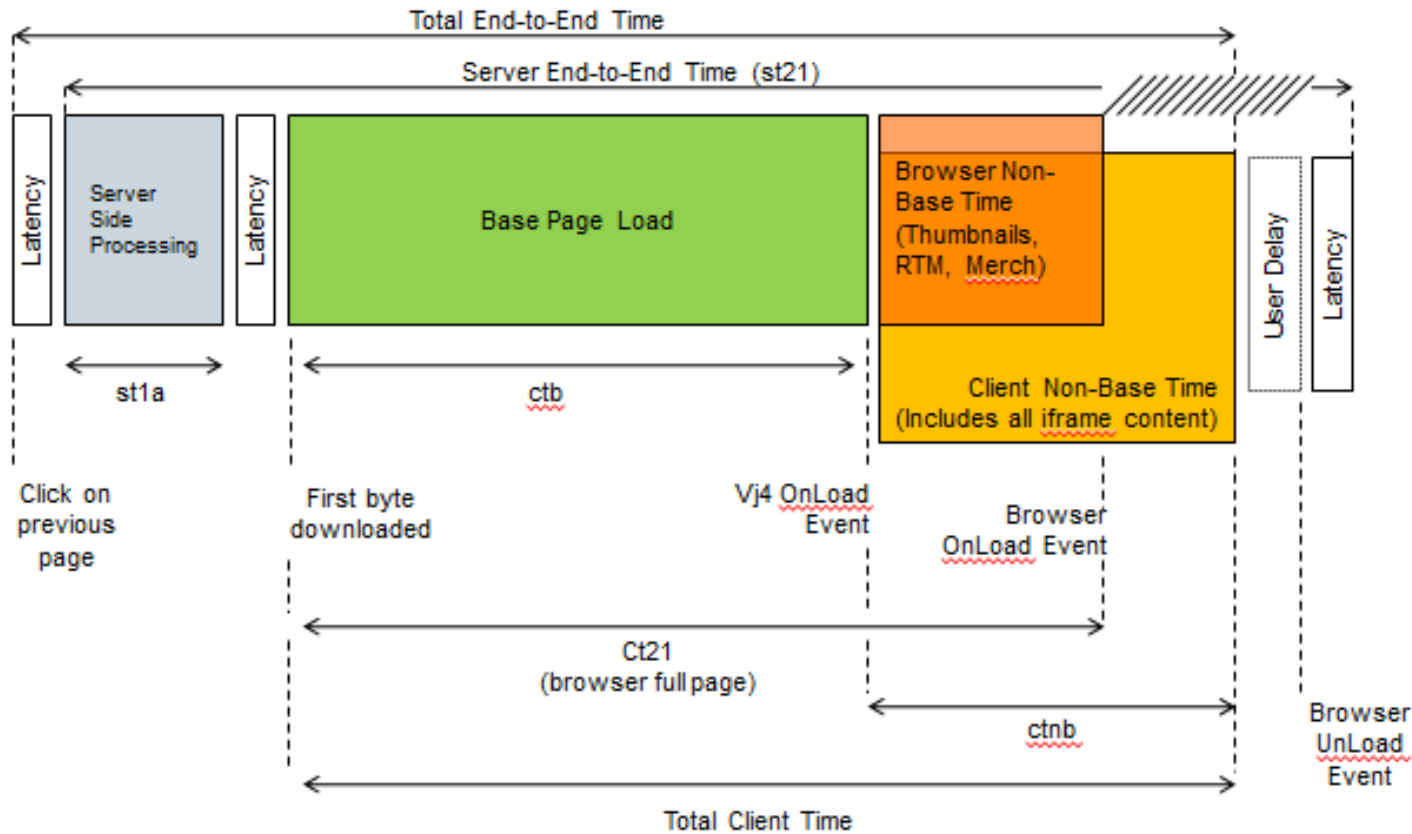
Topic started 1 week, 3 days ago by Zhong Li



R Case Study: How to Segment EBAY Mobile Buyers? (1 NEW)

Topic started 2 weeks, 3 days ago by Zhong Li

Site Speed Introduction



St1a – server side processing time

Ctb – base page load time

Ctnb – non-base time(thumbnails, RTM, Merch)

Site Speed Importance

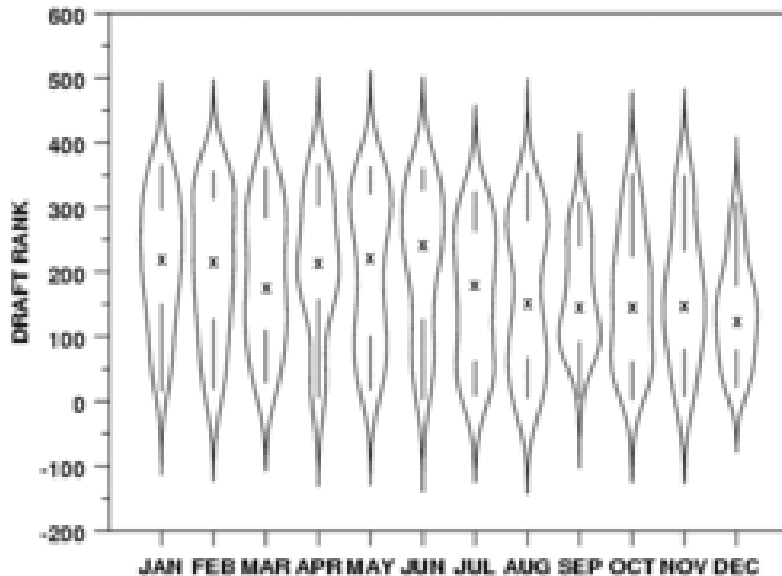
Load time is extremely important to Search Engine Results Page (SERP) and bottom line.

For example, a page load time of 8 seconds or more, according to the infographic, will result in approximately a 30% page abandonment.

47% of consumers expect 1 web page to load in 2 seconds or less,

If an e-commerce is making \$100,000 per day, a 1 second page delay could potentially cost you \$2.5 million in lost sales every year.

Violin Plot Introduction



http://en.wikipedia.org/wiki/Violin_plot

Violin plots are a method of plotting numeric data.

A violin plot is a combination of a [box plot](#) and a [kernel density plot](#).

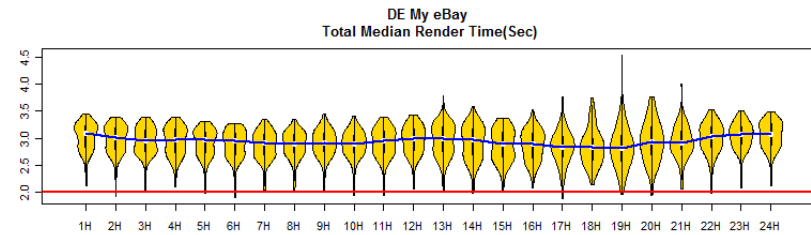
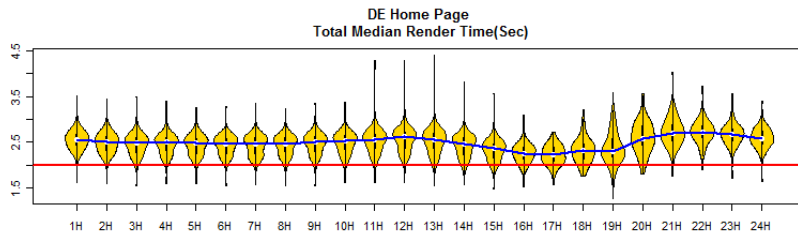
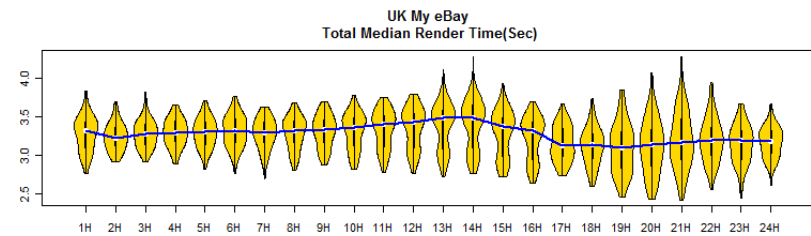
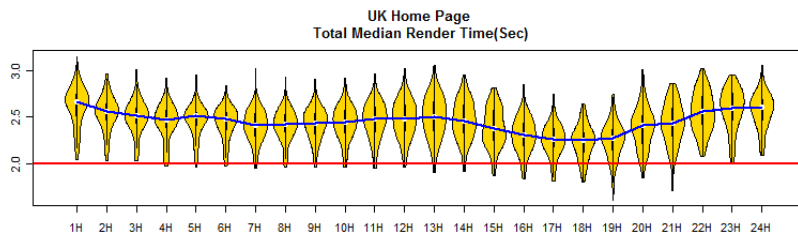
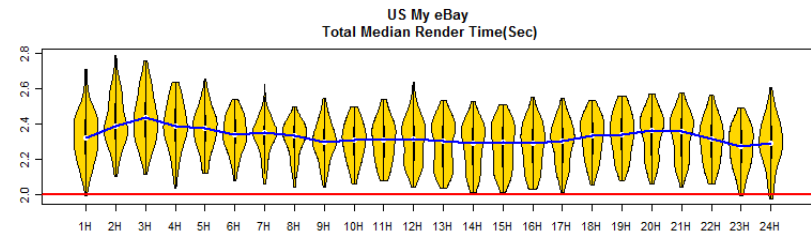
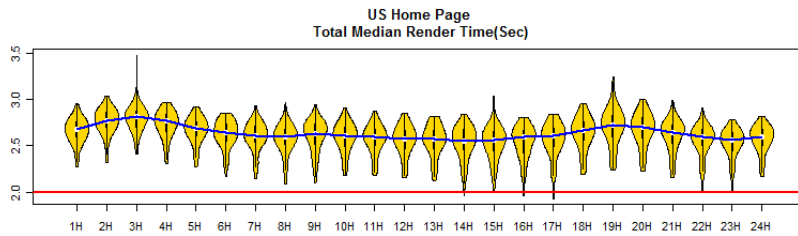
Specifically, it starts with a box plot. It then adds a rotated kernel density plot to each side of the box plot.

The violin plot is similar to [box plots](#), except that they also show the [probability density](#) of the data at different values (in the simplest case this could be a [histogram](#)).

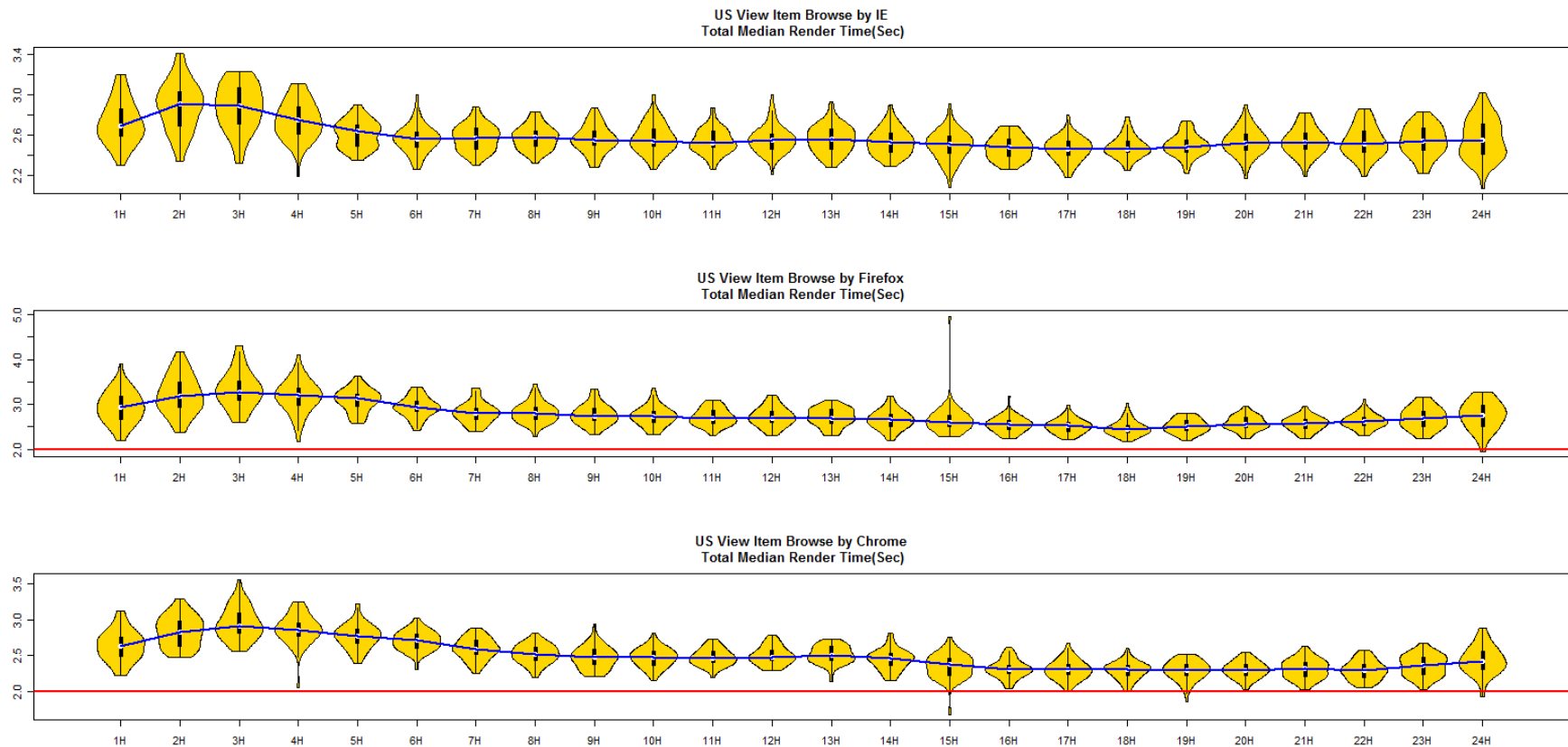
Typically violin plots will include a marker for the median of the data and a

box indicating the interquartile range, as in standard box plots. Overlaid on this box plot is a [kernel density estimation](#).

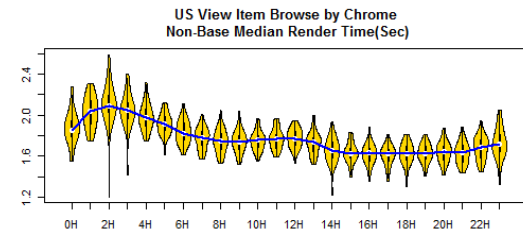
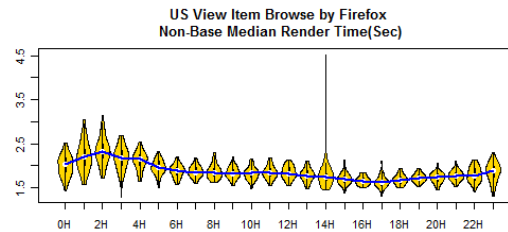
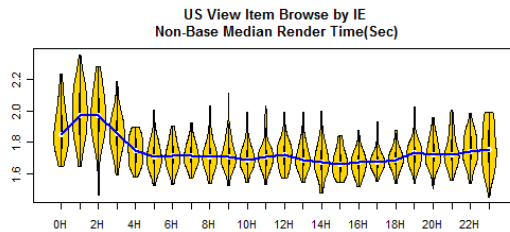
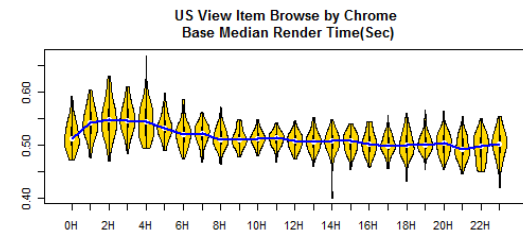
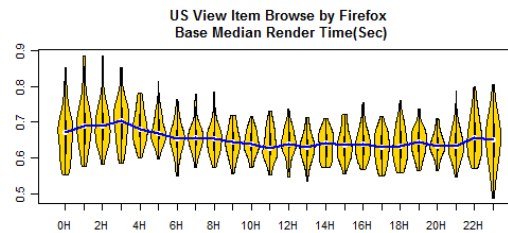
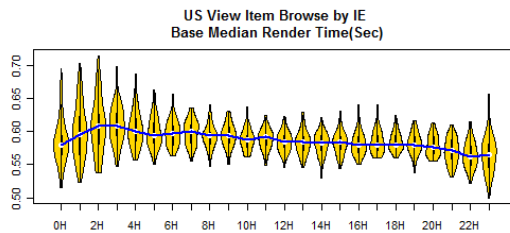
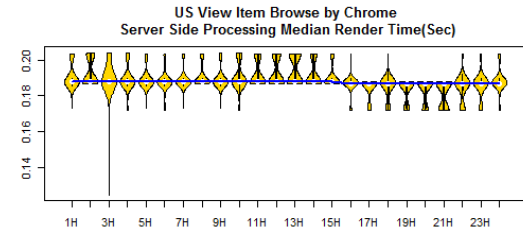
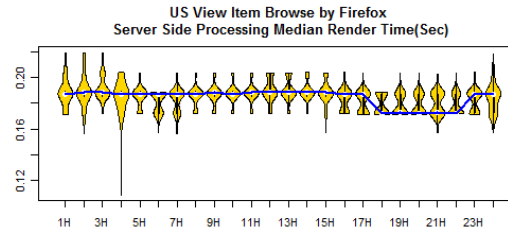
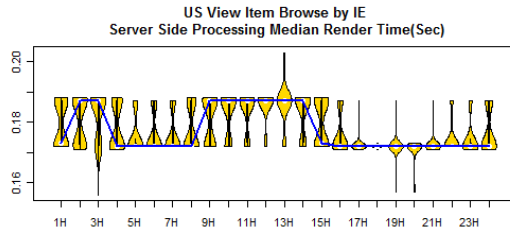
Page Level Site Speed Analysis



Browser Level Site Speed Analysis

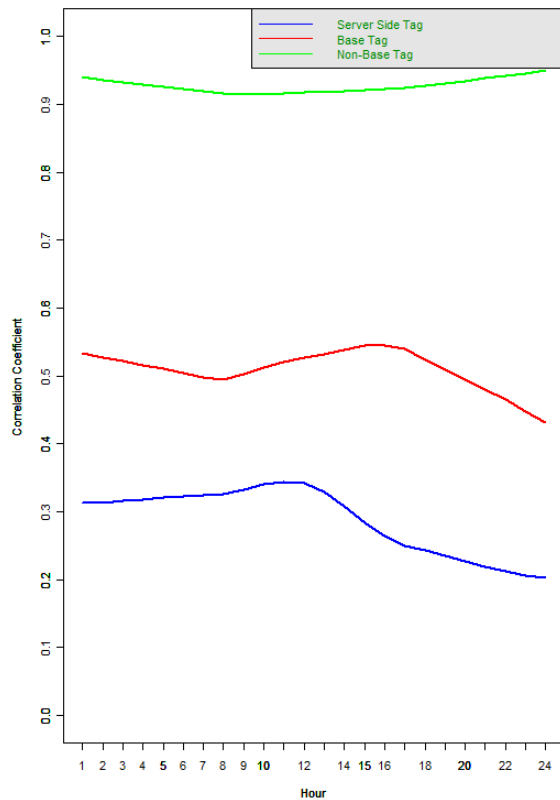


Tag Level Site Speed Analysis

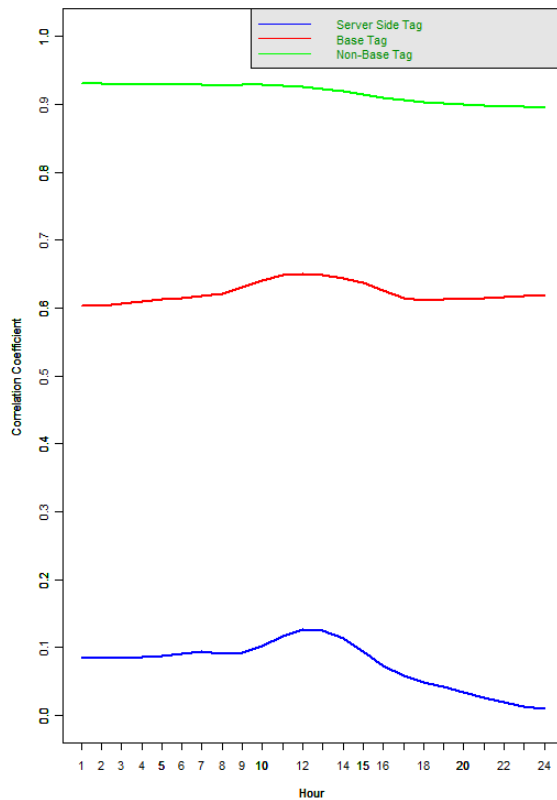


Page/Tag Correlation Matrix

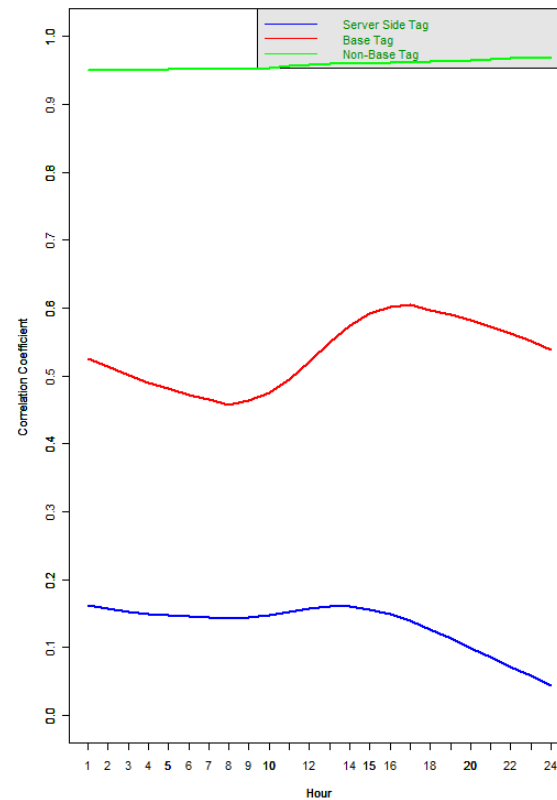
3 Key Page Tags Correlation Coefficient
for US View Item Browsed by IE



3 Key Page Tags Correlation Coefficient
for US View Item Browsed by Firefox



3 Key Page Tags Correlation Coefficient
for US View Item Browsed by Chrome



Site Speed Conclusion

- Different country has different peak and valley surfing time, can we have dynamic load-balancer to evenly distribute workload into site resource pools?
- Based on correlation matrix, the view item page is highly correlated to non-base tag, so if we want to improve site speed, we need consider cut down thumbnails, RTM, Merchandise.

Buyer Segmentation : What is RFM?

RFM is a method used for analyzing customer behavior and defining market segments.

RFM stands for

- **Recency** - *How recently did the customer purchase?*
- **Frequency** - *How often do they purchase?*
- **Monetary Value** - *How much do they spend?*

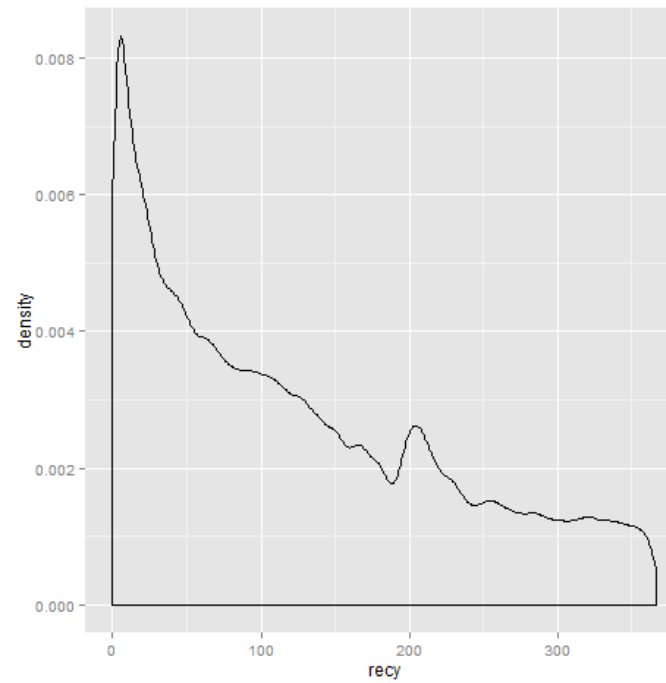
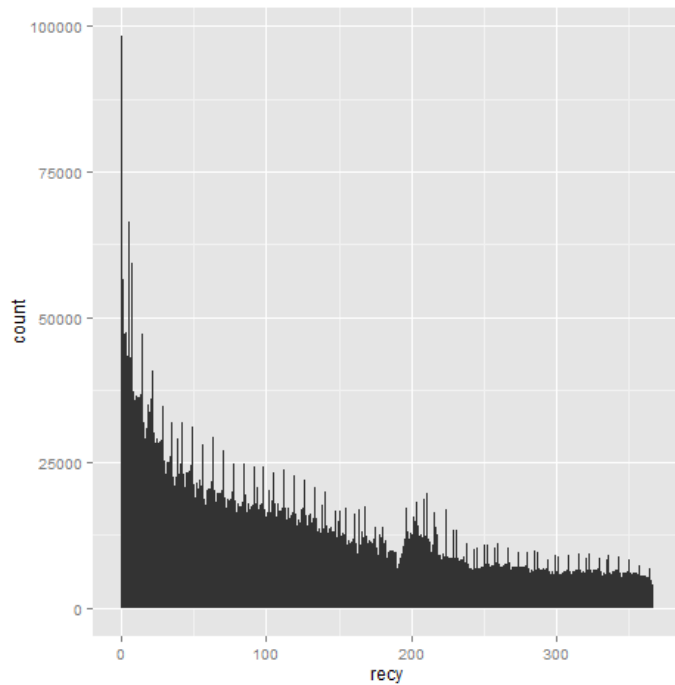
Buyer Segmentation : Data Preparation

Data preparation is the most important parts of the data mining process. In this step, the EBAY USA mobile buyer's purchase data for the previous 12 months will be converted to an appropriate format for the RFM model .

Three metrics that are used for RFM model will be calculated based on the below logic

- **Recency** - The interval between the latest purchase behavior happens and present.
- **Frequency**- The number of transactions that a customer has made within the last 12
- **Monetary**- The cumulative total of money(USD) spent by a particular customer.

Buyer Segmentation : RFM Metric Visualization



Buyer Segmentation : RFM Metric Normalization

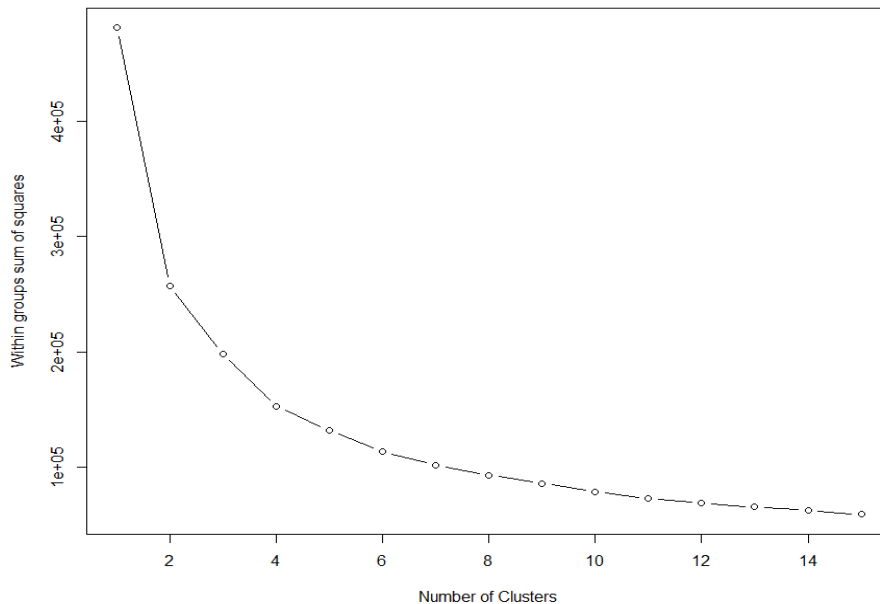
Due to skewed distribution for recency, frequency and monetary, I will consider to adopt log normalization which uses logarithms to better represent data that is highly skewed. Log normalization is helpful when values are clustered around small values with few large values.

Due to the correlation between recency and customer loyalty is negative, so the log_recy will be 1 minus logarithmic recency value.

- Normalized recency <- $1 - (\log(\text{recy} + 1) / \log(\max(\text{recy})))$
- Normalized frequency <- $(\log(\text{freq} + 1) / \log(\max(\text{freq})))$
- Normalized monetary <- $(\log(\text{monty} + 1) / \log(\max(\text{monty})))$

Cluster Number K Parameter Estimation

Considering the RFM 8 different variations ($2 \times 2 \times 2$) and easy marketing operation, I will choose 8 as the cluster number, also based on the below SSE(sum of squared error) graph, 8 is also the reasonable number for the model.



Buyer Segmentation : K-means++ Cluster

I will use k-means++(instead of k-means) algorithm to segment the mobile customer, here is the reason why I choose k-means++.

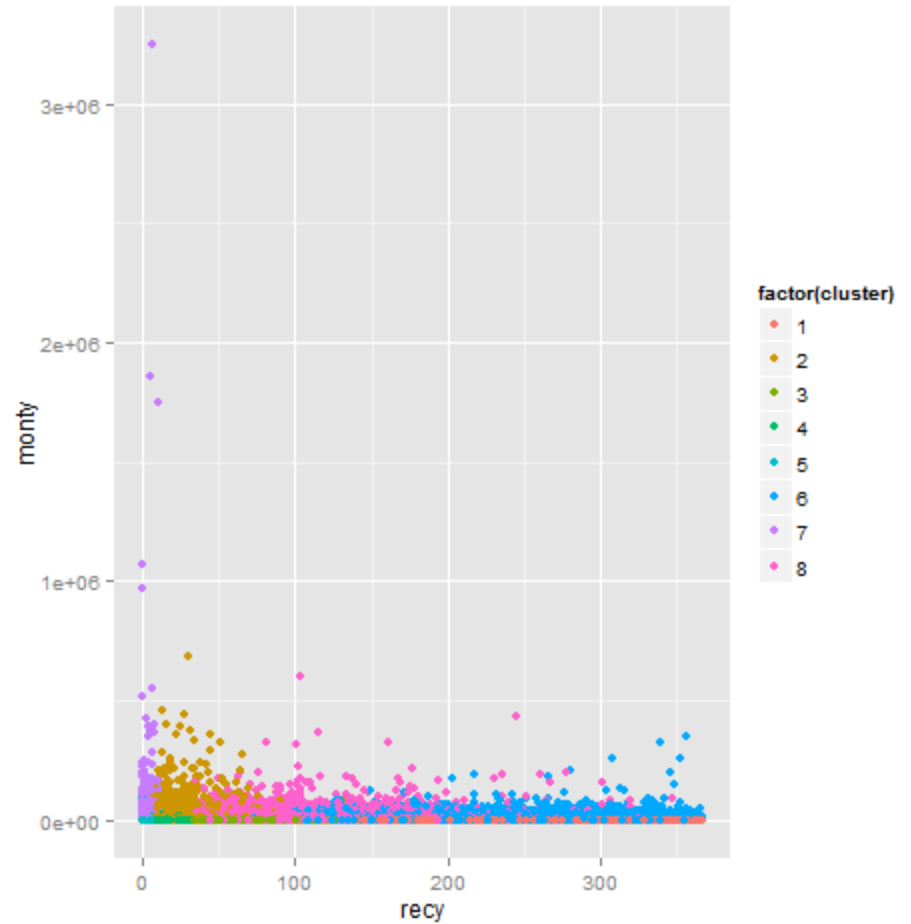
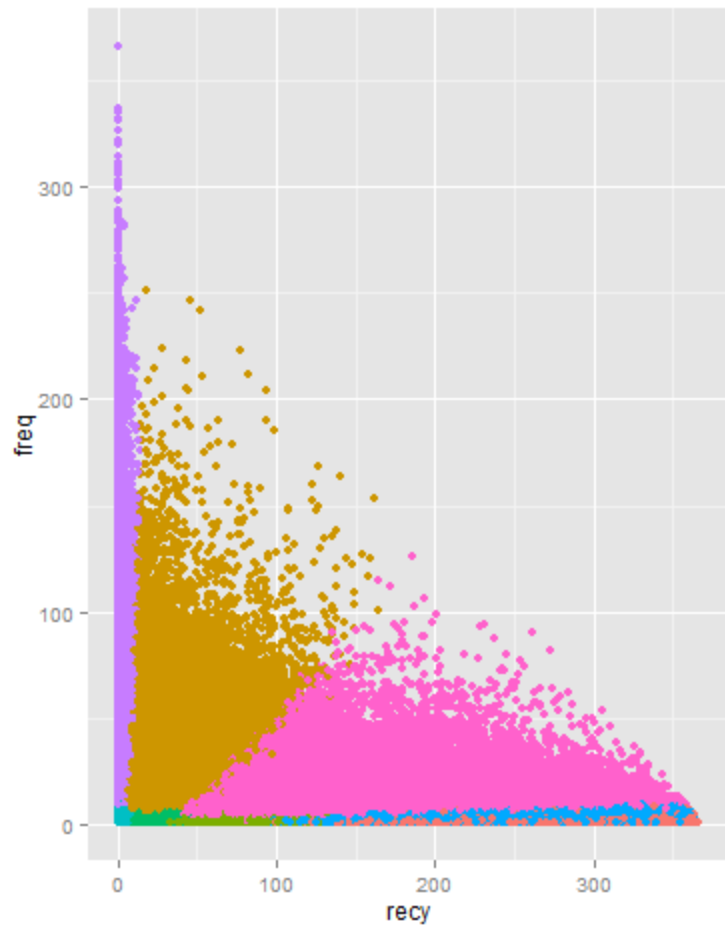
The *k*-means algorithm has at least two major theoretic shortcomings:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

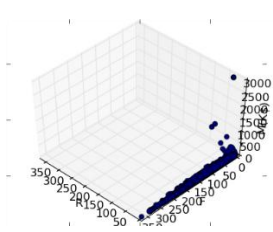
The *k*-means++ algorithm addresses the second of these obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard *k*-means optimization iterations.

With the *k*-means++ initialization, the algorithm is guaranteed to find a solution that is $O(\log k)$ competitive to the optimal *k*-means solution.

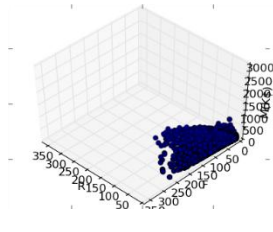
2D Customer Segmentation Visualization



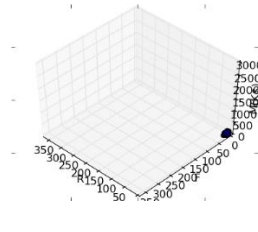
3D Customer Segmentation Visualization



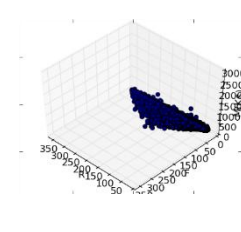
Best



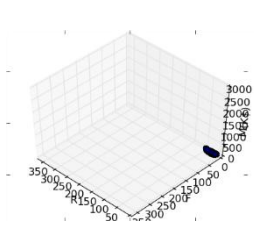
Valuable



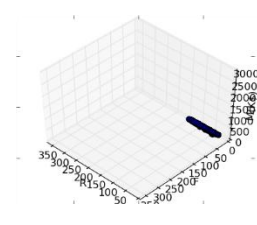
Shopper



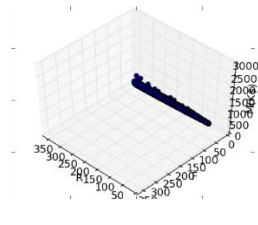
Churn



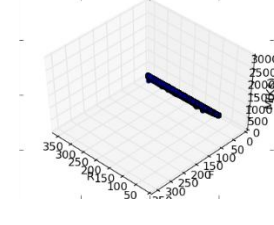
Recent Visitor



RV Churn



Spend



Uncertain

Case Study Summary

C7(Best, $R\uparrow F\uparrow M\uparrow$) - the most valuable customer segment, because it consists of customers who have recently made frequent purchases, and also have higher average purchase frequency and purchase amount.

C2(Valuable, $R\uparrow F\uparrow M\uparrow$) - the next valuable customer segment who has nearly the same characteristic as C7's customer.

C5(Shopper, $R\uparrow F\uparrow M\downarrow$) - the frequent shopper segment, they purchase frequently but with low monetary, so they are the best candidates for cross-sell and up-sell marketing campaigns.

C8(Churn, $R\downarrow F\uparrow M\uparrow$) - the churn segment, they have made a high number of purchases with high monetary values but not for a long time. It seems to be an indicator of churn likelihood. So we need contact with these customers by e-mail or to plan a customer reactivation program.

Case Study Summary

C4(Recent Visitor, $R\uparrow F\downarrow M\downarrow$)- the recent visitor segment, they have recently visited the EBAY site, with higher recency and lower purchase frequency and monetary value.

C3(Recent Visitor Churn, $R\uparrow F\downarrow M\downarrow$) - the recent visitor churn segment, they have visited the EBAY site not long time ago, with higher recency and lower purchase frequency and monetary value, but indicate will churn.

C6(Spender, $R\downarrow F\downarrow M\uparrow$) – the spender segment, they not visit EBAY site recently and frequently, but if they come to EBAY, they will purchase a lot.

C1(Uncertain, $R\downarrow F\downarrow M\downarrow$) - the least valuable segment for EBAY business, they are generally the least likely to buy again.

Lesson and Learn

- Ask the right business question.
- Fetch the different data points to generate various feature sets
- Build the test model based on small data set
- Aggregate different algorithms to improve the model quality
- Implement the production model based on Teradata or Hadoop
- Continuously improve the model



Thank You

Q & A

