Data Catalogs are Changing the Nature of Working with Data

Deb Seys

Strata Data Conference – May 2019

I'm here to talk to you about the results of some very interesting research that I had the opportunity to do while I was at eBay.

Research

- **Learning by Doing versus Learning by Viewing**: An Empirical Study of Data Analyst Productivity on a Collaborative Platform at eBay – presented at the 2018 ACM Computer-Supported Cooperative Work and Social Computing conference and published in the proceedings
  - https://www.kellogg.northwestern.edu/faculty/vanmieghem/htm/pubs/Learning-by-Viewing-at-eBay-cscw18_article193.pdf
- What's the Best Way to Learn a New Skill—by Doing or by Viewing? Kellogg 'Insights' newsletter 2019 – summary of the work and conclusions
  - https://insight.kellogg.northwestern.edu/article/best-way-learn-experience-curve
- Previous research at Kellogg into the value of collaboration among professionals in a hospital setting; Alation suggested looking at data analyst work in a similar way using data available from eBay.
  - Yue Yin, Kellogg; Itai Gurvich, Cornell; Jan A. Van Mieghem, Kellogg

Kellogg School of Management

ebay

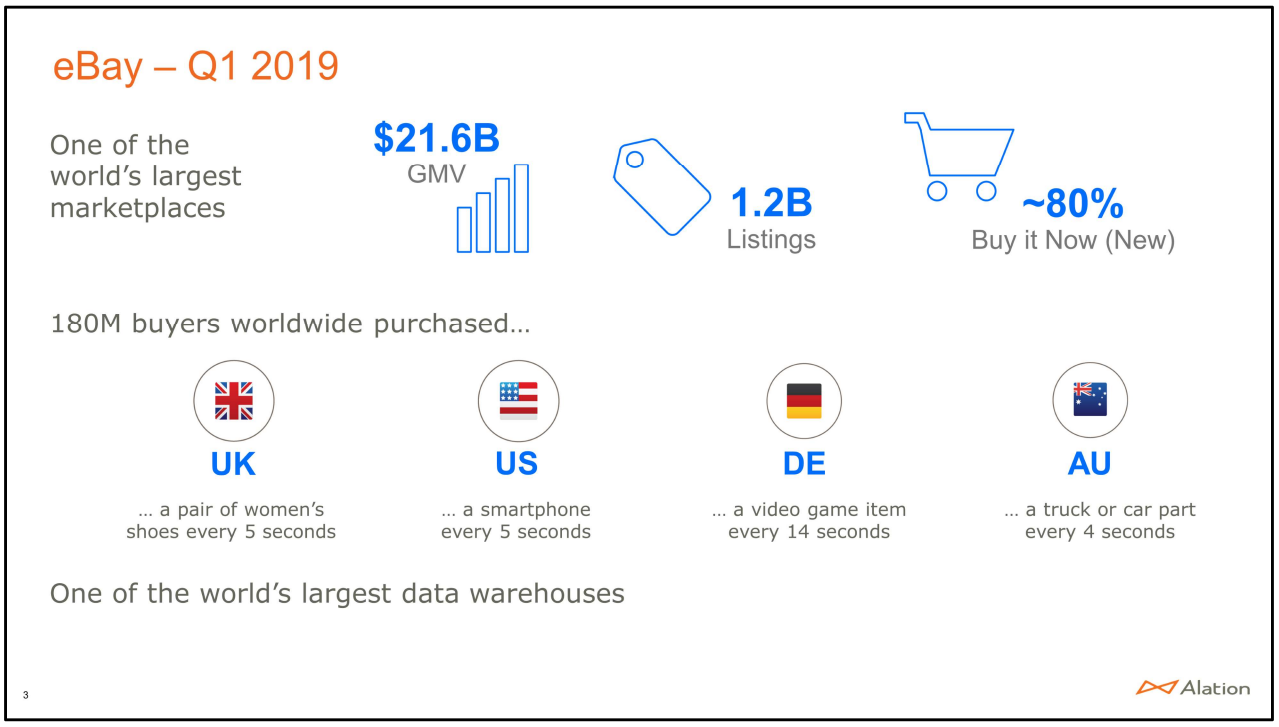
Alation

2

Research paper has the detailed analysis; charts, data, etc.

2016 started with in person visit and interviews with small number of analysts and data scientists; originally posed as an observational study of analyst work processes; but this too difficult to do in the eBay environment – analysts don't make easy shadow subjects and don't show up to work at the same time and place

After looking at usage data decided to focus on **query writing as a measure and a product of analyst work**



Context for eBay – scale, velocity, variety

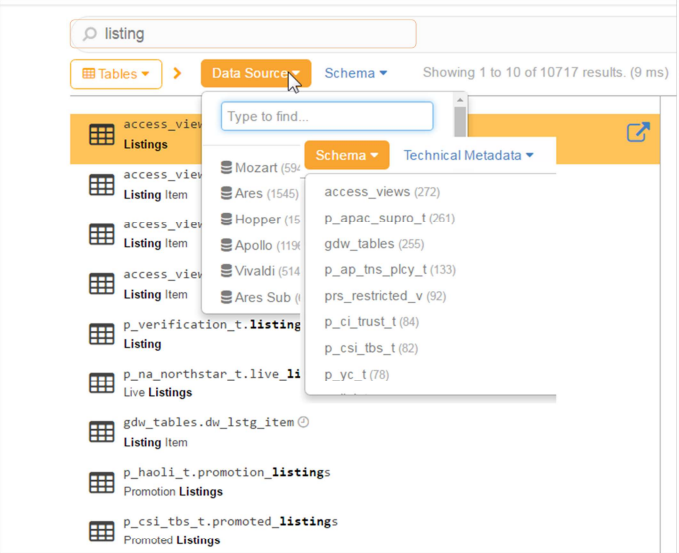
2015 – car bought via mobile app in the UK every 2 minutes.

Grand old timer of the ecommerce - 23 years of collecting data – organic and at internet speed

Nearly all of it available for analysis in the data warehouse

Discovery and Exploration

What data do I use?
How can I get the data?
How do I use the data correctly?



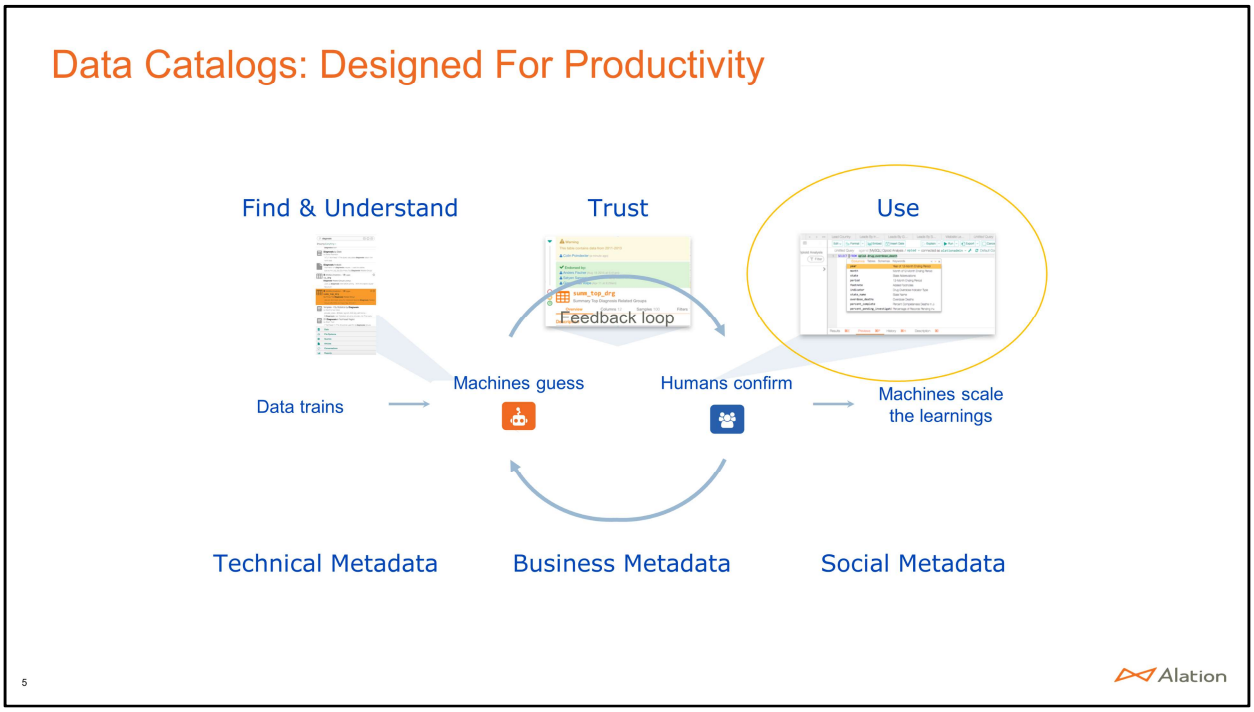
4

Alation

Over 10K tables that reference Listings data across ~4 major production systems in hundreds of schema

eBay began using a data catalog in 2014 to improve analyst productivity and to encourage collaboration and learning

- Faster data discovery and exploration
- Allow users to contribute
- Enable users to see each others work



Focus in on the Use phase

Data users need help along two different aspects
1) Familiar with the business domain and understand the data, but not proficient in query writing or working in a new query language
2) Proficient in querying, but unfamiliar with the data (which listings table has what I need?)

What we were looking at was whether the collaboration and sharing aspects of the use of Alation would result in measurable improvements in productivity

- Shortening the time spent in the discovery phase
- Building and maintaining proficiency in SQL
- Gaining and sustaining familiarity and expertise with the data itself

Data Set: eBay on Alation 2014 - 2018

4	2,001	79,797
Years	'Analysts'	Queries

- Alation usage data included
  - Records of all queries on Alation
  - Records of query executions on Alation
  - Records of all users viewing query pages on Alation
  - Records of all users on Alation
- 13% (1 out of 8) of all queries were viewed by analysts other than the authors

Alation

'Analyst' query writers could be data developers; development engineers, data scientists, product managers as well as typical data analysts

Original usage data much larger, this is after cleanup for users who were there the entire study period; not executives, not junior, not admins, etc. Cleanup of queries not executed, etc.

During study period - 10049 queries authored by 1097 data analysts had been viewed by analysts other than the authors

Over time that # continues to rise and I would expect to see the results magnified as this somewhat new and learned behavior becomes more common

Measuring Learning

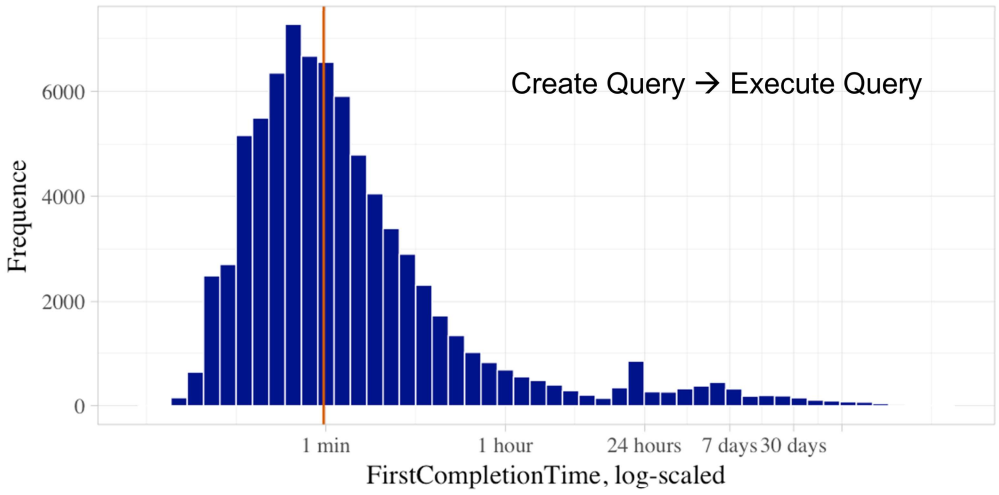
> The studies on programming have demonstrated that much of the **programming knowledge is tacit knowledge**, i.e. knowledge that is **not openly expressed or taught**.
>
> – SOLOWAY ET AL (1982), WAGNER & STERNBERG (1985)

So the study set out to measure learning and took some basis in studies that had been done in the programming world – as writing SQL (query writing) is a very similar activity and a performance change (therefore improved productivity) of data analysts can be associated with their learning.

Two kinds of measures - qualitative investigations like questionnaires, interviews and verbal protocol analyses – these are not good at capturing tacit knowledge. Or - practice or performance which have been shown to capture the tacit knowledge well.

This study used the behavioral approach of measuring performance change - so how do we do that?

Measure: First Completion Time



8

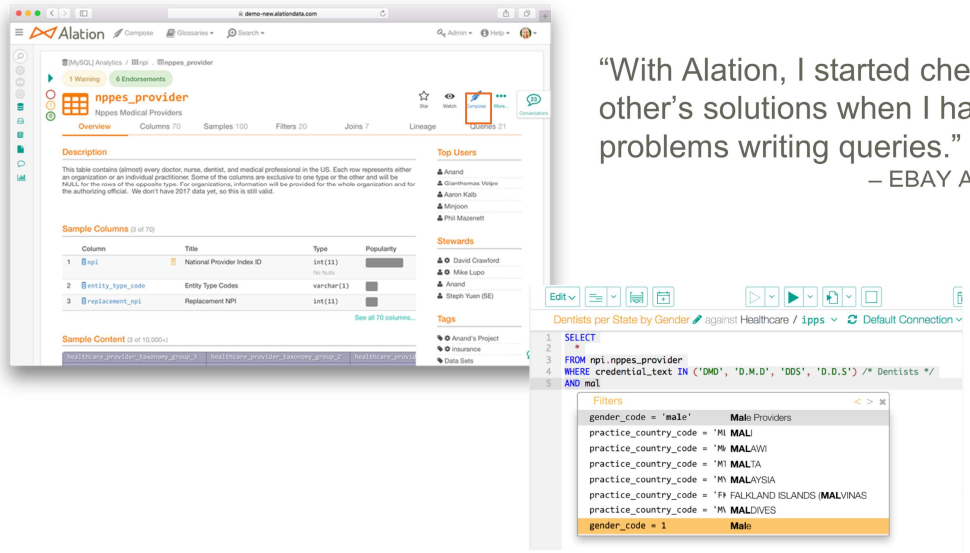
Alation

FirstCompletionTime used as a proxy for data analyst learning/productivity

This is the time between the moment when the analyst clicks create query (blank workspace) and runs the query - the data analyst shapes their idea into runnable query – the performance goal becomes the reduction of FirstCompletionTime

At that point we know a great deal
About the analysts
How many queries that analyst had previously written at that point; what databases they've queried
How many queries created by other analysts they had viewed at that point
How many users had viewed certain queries, etc

Learning By Viewing Vs. Learning By Doing



"With Alation, I started checking other's solutions when I have problems writing queries."

– EBAY ANALYST

9

Alation

So – for a single query writer whose FirstCompletionTime is going down…. Why?

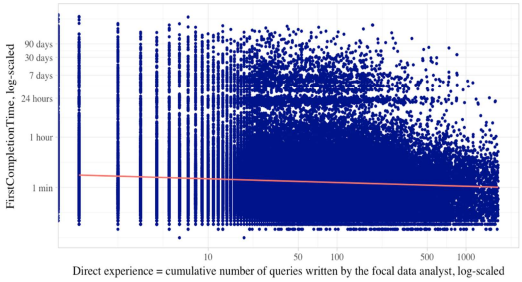
Learning by Doing = Direct Experience, writing own queries – using repetition and past experience

Learning by Viewing = Indirect Experience viewing queries written by peers – using Alation query library allows me to see others queries either by searching for a person or associated with data assets that I'm interested in

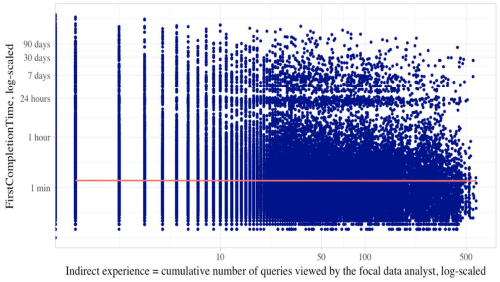
Note some of the background for this 'viewing' analysis comes from work done studying computer programmers; a learner's repertoire of programming concepts increases when engaged in remixing (reworking and combining existing creative artifacts)

Productivity Gains: Doing vs. Viewing

Learning by Doing
(9.1% increase in productivity*)



Learning by Viewing
(13-17% increase in productivity**)



* Increase associated with using a focal (familiar) database.

** Increase associated with viewing the work of certain analysts based on their output rate and social influence.

Alation

So of course, doing and viewing are both associated with productivity gains – BUT the data was variable and we wanted to know why – so it turns out

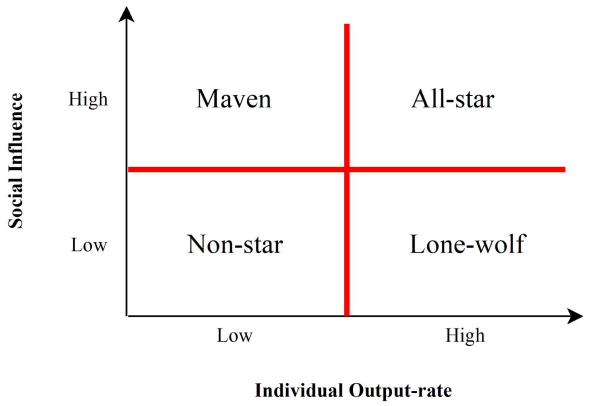
Doing – associated with significant productivity improvement when the analyst repetition is against the same database (moving from one to another db did not produce the same reduction in FirstCompletionTime)

Viewing – turns out that viewing queries that are authored by viewing the work of **certain** analysts is better.

Learning by viewing isn't necessarily more effective than learning by doing - rather, it depends who you're learning from…

Not All Query Authors Are Equal

- 4 Types of query authors segmented by output and social influence
- Output = the avg number of queries/month
- Social influence = social network analysis
  - Node = a data analyst
  - Edge A->B = analyst A views queries authored by B
- 2 ways to measure social influence
  - Viewership – local influence
  - PageRank – global influence



Alation

Segmentation along two axis - output and social influence

Social influence measured both by
Viewership (local) avg. num of distinct viewers per month of all queries authored by the Analyst
Page rank (global) based on social network (fewer direct viewers but those that do view are themselves highly viewed)

Lone-wolf very interesting here

Conclusions

- Learning by doing has a productivity increase when it is focused - 1.9 hours
  - Direct experience with the same database means 1.9 to 2.3 hours on average less completion time

Viewership Model = Lone-wolf best at 4 hours, but All-star also good at 3.3 hours

- Learning by viewing has a productivity increase under the viewership model:
  - 4 hours on average less for “Lone-wolf” queries
  - 3.3 hours on average less completion time for ”All-star” queries
  - Non-star & maven queries only increase query times

PageRank Model = All-star Queries best at 4.3 hours

- Learning by viewing has a productivity increase under the PageRank model:
  - 4.3 hours on average less completion time for ”All-star” queries
  - 4 hours on average less for “Lone-wolf” queries
  - Non-star & maven queries only increase query times

Alation

Limitations and Suggestions for Future Investigation

- Measures of productivity and learning
  - More investigation needed into “first completion time” as measure of productivity
  - May be related to coding styles (such as tinkerers or planners).
  - What exactly does a data user learn that affects first completion time?
  - This study is quantitative only - more attention to quality or success of results of the query itself.
- Social influence
  - More research identified on Mavens versus Lone Wolf
  - Why is social influence a factor in learning?
- Collaborative platform
  - Has measurable benefits
  - Needs ease of capture of activities and ease of discovery for viewing by others
- Organizational resourcing
  - Value mastery over the data domain as well as technical expertise
  - Need to incentivize “All-star” analysts and their work

13

Alation

Treat the increase of learning as a more apprentice-like process than as if one were training fungible human resources with easily transferable skills

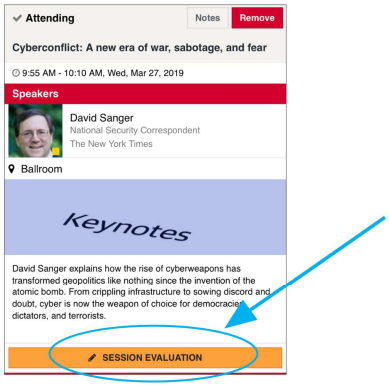
Learn more…

- Visit the Alation Booth #210
- Find out why data culture might make you pack a space suit
- Get scanned and receive a copy of this presentation and cool socks!

Alation

Rate today's session

Session page on conference website



O'Reilly events app

Alation


Alation