Copyright © 2016 Splunk Inc.

A Mathematical Perspective on Data Science

Dr. Tom LaGatta Staff Sales Engineer (previously Staff Data Scientist)

28.251.81 - - [02/Feb/2014 Screen?cstegory_idsTEDW+462

category_id=FLOWERS' Mozilla/401 ETEDDY&JSESSIONDESDSLAFTAGEN ategory_id=TEDDY+ Mozilla/40 (company) id=GIFTS8.JSESSIONO-SD15L6FF3AUF804, No. 2gory_id=GIFTS+ Mozilla/5.0 (Windows I); tex SONID=SD4SL5FF2ADFF1 HTTP 11*40 200 STATE RP-LI-02* Googlebot/21 (http://www.gooddate reen?category_id=FLOWERS&JEESMOn-Seemon zilla/5.0 (Macintosh; U; Intel Nac 05 % 10.6.3 extinguisting) GET /product.screen?product_id=1-Sit-6id.example.sc temld=EST-16&product_id=FI-5W-dt+Weak50 (Koldet Nation 228,251,81 - - [02/Feb/2011:16:0023] & (point-main/information)

y.screen?category_id=TEDDY* MozilaSU (Namet, National Contemport 62 - - [02/Feb/2011:16:00:23] @7 /protectanos/point integ category_id=FLOWERS' Mozilla/4.0 (compatible NE 66, No.00 d=TEDDY8: JSESSIONID=SD9SL4FF4A0FF8 HTP 12:00 383 find

ategory_id=TEDDY* Mozilla/4.0 (competible; HEE (); Buier R id=GIFTS&JSESSIOND=SD1SL6FF3A0F78 HTP 12 (# 322 104) egory_id=GIFTS* Mozilla/5.0 (Windows) II; Windows (II) SIONID=SD4SL5FF2ADFF1 HTTP 11 400 2012 @7.4 -RP-LI-02* Googlebot/21 (http://mmga

reen?category_id=FLOWERS&RESSID_0080 rilla/5.0 (Macintosh; U; Intel Nac 06 110 C304) GET /product.screen?product.id=A.SPL40

temId=EST-168:product_id=FI-SW-01 (Note) 228.251.81 - - [02/Feb/201116-0023] @

1.screen?category_id=TEDDP Hoods50 IN 62 - - [02/Feb/2011.16.0023] @1/P ategory_id=FLOWERS' Noviet (1) H=TEDDY&JSESSIONID=SDSSL4F4407

tegory_id=TEDOY' Mozilla/4.0 (co CIETSS: ISESSIONID-SDISL OFFICE



About Me

- Math PhD from University of Arizona
 - "Geodesics of Random Riemannian Metrics" w/ Janek Wehr
 - Probability + Differential Geometry + Functional Analysis
- Postdoc at Courant Institute @ NYU
 - Finished Geodesics paper, published in Communications in Math. Physics
 - Collaborated with Political Scientists on heterogeneous voting behavior
- Was: Staff Data Scientist at Splunk
 - Helped customers with advanced use cases in Business Analytics, Internet of Things, Machine Learning, Data Science
- Now: Staff Sales Engineer at Splunk
 - Helping big big customers solve big big business problems

62 – – LU2/Feb/2011:16:00:23] GET /product.screen?product_id=FFW-02.05300-9456660333 Statemanteernations category_id=FLOWERS* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 51; SN; NT 01102333 Working and the product of the d=TEDDY8: JSESSIONID=SD9SL4FF4ADFF8 HTTP 1.1*200 3439 Windows NT 51; SN; NT 0110234040033041 WT (colores produc ategory_id=TEDDY* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 51; SN; NT 011023404003041 WT (colores product)



Abstract

As with all things, the process of analyzing data admits a mathematical description. As a mathematician-turned-data-scientist, I will describe my approach to problem solving, and attempt to loosely formalize "stakeholders", "use cases", "data" and "deliverables" in mathematical language for the enjoyment of this mostly academic audience. In particular, I will describe how query languages are inherently functional, acting as functional transformations of Data into Data, which obey the usual functional composition law. The process of analyzing data results in an iterative sequence of queries, converging to a final query which is satisfactory to the use case. These queries are then organized into deliverables, which can be "dashboards" (web pages with visualizations) or "data products" (with scheduled jobs & analyses running in the background). When this process is done right, it results in the extraction of "value" for stakeholders, which can be measured tangibly in terms of revenue, costs or risk metrics. Sometimes this has a fancy name like "data science", but more often than not, is just the normal operational work of a good datasavvy IT, Security, Tech or Business department in modern enterprises and governmental agencies. There will be no proofs, but I would be very interested to discuss rigorous approaches to social organization & problem solving after the talk.



Agenda

- Basic Definitions:
 - Data, Stakeholders, Use Cases, Deliverables
- Query languages as functional programming
 - Every query is a map f: Data -> Data
 - Example problem solving
- Putting It All Together: Doing Data Science
 - Emphasize actionable insights
 - Tie it together to deliver "value" to stakeholders



Copyright © 2016 Splunk Inc.

Basic Definitions

-0.251.81 - - [02/Feb/2011.10 4.5creen?category_ide1E0/201106023[10] S2 - - [02/Feb/2011160023]10] S2 - - [02/Feb/2011160023]10]

sategory_id=FLOWERS' Mozille/ I for ategory_id=TEDDY+ MozillalAD (competition id=GIFTS8.JSESSIONO-SD15L6FF3AUF804, No. egory_id=GIFTS+ Mozilla/5.0 (Window; 1, Wei SONID=SD4SL5FF2ADFF1 HTTP 13*40 200 STA RP-LI-02* Googlebot/21 (http://www.goodade reen?category_id=FLOWERS&XESSION=Seams zilla/5.0 (Macintosh; U; Intel Nac 05 % 10.6.3 extinguisting) GET /product.screen?product_id=1-Sit-464.game.equ temld=EST-16&product_id=FI-5W-dt+Weak50 (Koldet Nation 228.251.81 - - [02/Feb/201116:0023] & Implementations y.screen?category_id=TEDDY* Nozila50 (kainek;kialmateory) mate 62 - - [02/Feb/2011:16:00:23] &T /productsmediprint illigarian

category_id=FLOWERS* Mozilla/4.0 (corpside; NEA, Solars, ST d=TEDDY8:JSESSIONID=SD9SL4FF4ADFF8 HTTP 11:00 300 Fabrics ategory_id=TEDDY" Mozilla/4.0 (compatible; NEE 0; Timine #A Id=GIFTS& JSESSIONID=SD1SL6FF34UFF3 HTP 114 32 Haddaal (egory_id=GIFTS* Mozilla/5.0 (Windows; U; Wider (15) + SIONID=SD4SL5FF2ADFF1 HTTP 11:400 2012 481 Hatu -RP-LI-02* Googlebot/2.1 (http://mm.gov reen?category_id=FLOWERS0_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSI rilla/5.0 (Macintosh; U; Intel Nac (6 1 11 1.5 and 14) GET /product.screen?product.id=F1.98.080 temld=EST-16&product_id=FI-SH-40 million 228.251.81 - - [02/Feb/2011364083] 67) LSCreen?category_id=TEOU* Notific# 62 - - [02/Feb/2011.16:0023] 67 /p ategory_id=FLOWERS' Nozilia/40 H=TEDDY& JSESSIOND=SD9SL4F1480F1

tegory_id=TEDDY Mozilla/40 (cont CIETSS ISESSIONDESDISLOF



What is Data?

- "Data" is any informational artifact of real-world phenomena
- A "metric" or "KPI" is any aggregate function of low-level data
- Examples:
 - Semi-structured timestamped events/metrics
 - Structured relational data (rows & columns)
 - Graph data (nodes & edges)
 - "Unstructured" data (images, video, text)
- How to model data:
 - Events: marked point processes & time series (Skorokhod space)
 - Relational schema: categories (see David Spivak's work)
 - Other data: depends on the use case, might need new data structures to represent it (incl. vectors, graphs, etc.)

2 – _ [02/Feb/2011:16:00:23] 'GET /product_screen?product_id=FiFW_028233333333334 Methylama.phase.asta ategory_id=FL0WERS* 'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 51; 91; JE011023331 Methylama.phase.asta = TEDDY8_JSESSIONID=SD9SL4FF4ADFF8 HTTP 1.1*200 3439 Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01103; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01104; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01104; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01104; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01104; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE01104; JE02104; Category Compatible; JE02104; Category Compatible; MSIE 6.0; Windows NT 51; 91; JE0210; JE

199.70.25.81 - - [10/May/2016 22:18:23:078770] "GET /oldlink?item_id=MCF-3&J SESSIONID=SD4SCL2FF6ADFF2 HTTP 1.1" 404 948 "http://shop.acme.com/cart.do?ac tion=remove&item_id=MCF-3&product_id=MCF-3" "mozilla/5.0 (Linux; U; Android 2.3.4; en-us; SonyEricssonLT15i Build/4.0.2.A.0.42) AppleWebKit/533.1 (KHTM L, like Gecko) Version/4.0 Mobile Safari/533.1" 3683



What is a Stakeholder?

- A "stakeholder" is a person, group or organization who is invested in the outcome of an initiative.
- Example: A company buys software
 - Stakeholder orgs are IT & the Business
 - Individual stakeholders include Individual
 Contributors, Managers, Directors & Executives.



- IT stakeholders have performance metrics (num. outages, mean time to resolution, etc)
- Business stakeholders have different metrics (revenue, cost, risk)
- Customer stakeholders downstream also have value/impact metrics

62 – – LU2/Feb/2011:16:00:23] 'GET /product.screen?product_id=FFW-026555556945556664555566455556 :ategory_id=FLOWERS' Mozilla/4.0 (compatible; MSIE 6.0; Windows III 51; SVI; ETOL10263351 Wei/www.spikers/ag.cor/core d=TEDDY&JSESSIONID=SD9SL4FF4ADFF8 HTTP 1.1*200 3439 Windows III 51; SVI; ETOL10264564665666646464646464646464 become id=TEDDY: Mozilla/4.0 (compatible; MSIE 6.0; Windows III 51; SVI; ETOL102043164646666666666666666666666



What is a Stakeholder? (cont.)

- How to model stakeholders?
 - Follow Game Theory for inspiration (but don't worry about "equilibrium")
 - Create an index set I with all stakeholders. Various actions, outcomes & objectives will have subscripts i based on stakeholders
 - E.g., person i chooses action a_{i,t} at time t
 - I can be hierarchical (Person i contained in org A, so parent(i) = A)
 - "Value" modeled by objective functions (U_{i,n} = objective #n for person i)
 - Is the action "pivotal" for the outcome? (ie $\mathbb{E}[U|do(a)] > \mathbb{E}[U|do(not a)]$?)
- Keep track of stakeholders data:
 - Might be high-level (email, Powerpoints) context is key
 - Might be in databases (transactions, customer records, tickets data)
 - Might be granular events data (web clickstream, logs, mobile, wire data)



What is a Use Case?

- A "use case" consists of a business problem, a strategy to alleviate the problem, metrics to evaluate the outcome, data to power a solution, and stakeholders who are involved in its development.
- Use Case: Problem Forecasting.
 - Company has costly network/system outages
 - IT hires a Data Scientist to help build solution.
 - Data includes Infrastructure (CPU, Memory),
 Operations (Outage Reports), App logs, etc
 - Metric: cost of outages ≈
 num outages * time to resolution * cost of labor
 - Stakeholders include IT & Business, and impacted customers

62 – – [U2/Feb/2011:16:00:23] 'GET /product.screen?product_id=N=W=02453300=905Head(now http://www.phone.sto category_id=FLOWERS* 'Mozilla/4.0 (compatible; MSIE 6.0; Windows HT 51; SY); ET 0111687831 MHp//www.phone.sto d=TEDDY&JSESSIONID=SD9SL4FF4ADFF8 HTTP 1.1*200 3439 Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400abaccory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400abaccory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011668/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011688/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011688/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 011688/HMMLBAAD(ST 14400-Stategory_id=TEDDY: 'Mozilla/4.0 (compatible: MSIE 6.0; Windows HT 51; SY); ET 0116888





What is a Deliverable?

- A "deliverable" is a thing produced to solve a use case
 - Can include "dashboards": informational web pages built from data queries
 - Or "workflows": notable events deliberated to operations analysts
 - Or full-fledged "data product": application which *does* stuff automatically
- Deliverable: Problem Forecasting System
 - Goal: forecast problems before they start, deliver "proximate root cause" to IT to investigate
 - Data: CPU, Memory, Latency, Service Tickets
 - Build machine learning model to correlate
 Infrastructure data with Service impact
 - Apply model to incoming events, create "predicted(Risk_Score)"
 - Surface high-risk events to IT Operations

eve	erity		Status		Name							
С	RITICAL	L 0	All	0 -			Job		🔲 🥊 Smart N	/ode >		
нідн 0								15 events (5/9/16 11:00:00.000 PM to 5/10/16 11:18:48.000 PM)				
		All		Θ.								
		0	Service				rmat Timelin Zoom to Sele		Zoom Out			
			× All							ir per c		
IN	IFO	0				16				. p		
					Last 24 hours 🗸 Submi	t 10						
							12:00 AM		12:00 PM			
							Tue May 1	0				
Edi f		lected Edit all 15 matchi	ing events Service(s)	Title			Severity	Status	3 Owner	Ac		
					d for kpi=DB Hosts CPU Load Percent of	service=DB	Severity Medium	Status	o Owner unassigned	Ac		
í		Time 5/10/16 11:18:44.000 PM 5/10/16 11:18:04.000 PM	Service(s) DB	Anomaly detected Service	d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of			New		~		
i >		Time 5/10/16 11:18:44.000 PM 5/10/16 11:18:04.000 PM 5/10/16 11:17:54.000 PM	Service(s) DB Service DB Service DB Service	Anomaly detected Service Anomaly detected Service Anomaly detected Service	d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of	service=DB service=DB	Medium	New New New	unassigned Open DB Servic Deep Dive Go to DB Servic	e Kpis		
í > >		Time 5/10/16 11:18:44.000 PM 5/10/16 11:18:04.000 PM 5/10/16 11:17:54.000	Service(s) DB Service DB Service DB	Anomaly detected Service Anomaly detected Service Anomaly detected Service	d for kpi=DB Hosts CPU Load Percent of	service=DB service=DB	Medium Medium	New New New	unassigned Open DB Servic Deep Dive	e Kpi e Las stigat		
i > > >		Time 5/10/16 11:18:44.000 PM 5/10/16 11:18:04.000 PM 5/10/16 11:17:54.000 PM 5/10/16 11:17:54.000	Service(s) DB Service DB Service DB Service DB	Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service	d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of	service=DB service=DB service=DB service=DB	Medium Medium Medium	New New New New	unassigned Open DB Servic Deep Dive Go to DB Servic Deep Dive Inves Open Drilldown Deep Dive	e Kpi e Las stigat Sear		
i > > >		Time \$/10/16 11:18:44.000 PM \$/10/16 11:18:04.000 PM \$/10/16 11:17:54.000 PM \$/10/16 11:17:54.000 PM \$/10/16 11:17:54.000 PM \$/10/16 11:17:41.000 PM \$/10/16 11:17:41.000 PM \$/10/16 11:17:41.000 PM	Service(s) DB Service DB Service DB Service DB Service DB	Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service	d for kpi-DB Hosts CPU Load Percent of d for kpi-DB Hosts CPU Load Percent of	service=DB service=DB service=DB service=DB service=DB	Medium Medium Medium Medium	New New New New	unassigned Open DB Servic Deep Dive Go to DB Servic Deep Dive Inves Open Drilldown	e Kpi e Las stigat Sear		
i > > >		Time 5/10/16 11:18:44.000 PM 5/10/16 11:18:04.000 PM 5/10/16 11:17:54.000 PM 5/10/16 11:17:54.000 PM 5/10/16 11:17:41.000 PM	Service(s) DB Service DB Service DB Service DB Service DB Service DB Service DB	Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service Anomaly detected Service	d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of d for kpi=DB Hosts CPU Load Percent of	service=DB service=DB service=DB service=DB service=DB service=DB	Medium Medium Medium Medium Medium	New New New New New	unassigned Open DB Servic Deep Dive Go to DB Servic Deep Dive Inves Open Drilldown Deep Dive Open Correlatio	e Kpi e Las stigat Sear n Sea		



Copyright © 2016 Splunk Inc.

Queries as Functional Programming

-0.251.81 - - [02/Feb/2011] Lon.od.od. J.Screen?category_id=[C00+Ko022] (014) S2 - - [02/Feb/2011:16:0023] (014) C011:16:0023] (014)

tategory_id=FLOWERS* Mozile/4.0 (no ETEDDY&JSESSIONDESDSLAFT4ADRAM ategory_id=TEDDY- Mozills/A0 (compatible) id=GIFTS8.JSESSIONO=S015L6F5AUF801 egory_id=GIFTS+ Mozilla/5.0 (Window; 1, Wei SONID=SD4SL5FF2ADFF1 HTTP 11*40 20 CT 10 RP-LI-02* Googlebot/21 (http://www.goolda.or reen?category_id=FLOWERS&RESAME-Signa zilla/5.0 (Macintosh; U; Intel Nac 06 % 10 £3 e40 (plan 'GET /product.screen?product.jd+1-58-468.8200-3006 temld=EST-168:product_id=FI-SW-01*Hookes()(vides to a state of the sta 228.251.81 - - [02/Feb/201116:0023] & Implementations

y.screen?category_id=TEDDY* Nozila50 (kainsk kialitation) 62 - - [02/Feb/2011:16:00:23] @T /productsmatignedic importan category_id=FLOWERS' Mozilla/4.0 (compatible NE 64, Tables, 5, 5 d=TEDDY8:JSESSIONID=SD9SL4FF4ADF78 http:///www.

ategory_id=TEDDY* Mozilla/4.0 (compatible; NEE 6); Table (PA id=GIFTS&JSESSIONID=SD1SL6FF3AUF78 MTP 12 # 32 https://doi.org/ egory_id=GIFTS' Mozilla/5.0 (Windows; I); Tinder II (Style SIONID=SD4SL5FF2ADFF1 HTTP 11 400 288 45 late -RP-LI-02* Googlebot/21 (http://mm.gov reen?category_id=FLOWERS0_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSIOD_SESSI rilla/5.0 (Macintosh; U; Intel Nac (6 8 10,6 and 100) GET /product.screen?product.id=Fl.sp.adu temld=EST-16&product_id=FFSW-UP https://www. 228.251.81 - - [02/Feb/2011160033] & 1.screen?category_id=TED0Y_Nod850.04 62 - - [02/Feb/2011.16.0023] @1/p ategory_id=FLOWERS' Mozilla/40 H=TEDDY& JSESSIONID=SD9SL4FF4NFF

tegory_id=TEODY' Nozilla/4.0 (con CUETSP. ISESSIONDESDISLOFFSM



Query Languages

- Query languages provide a formulaic approach to working with data
- A "query" is a string which tells where to get the data, what to do with the data, and where to put the data (incl visualization or DB)
- Mathematically, every query is a FUNCTION f : Data -> Data
- Queries can be composed (with | symbol), and analysis is iterative
- Example 1: Successful Purchase Actions from Web Logs sourcetype=access_combined action=purchase status=200
- Example 2: Plot Purchase Value as Metric Timeseries

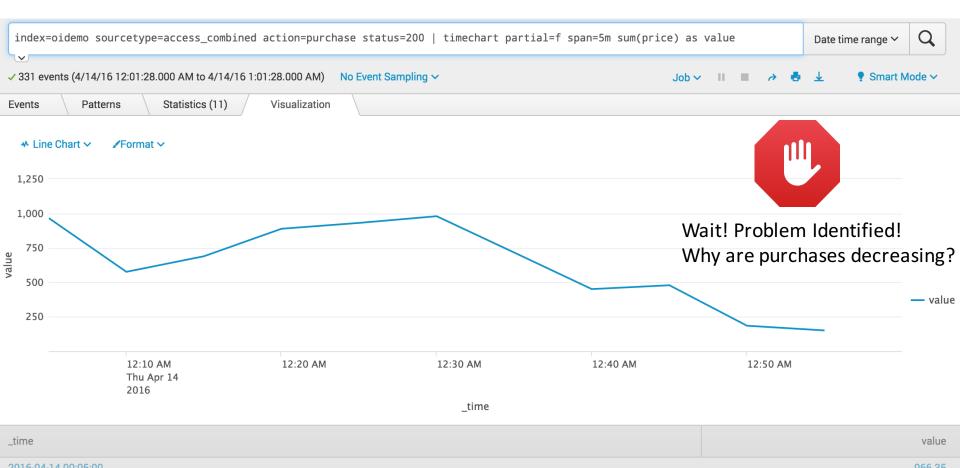
sourcetype=access_combined action=purchase status=200
| timechart partial=f span=5m sum(price) as value



1: Successful Purchase Actions from Web Logs

<pre>index=oidemo sourcetype=access_combined action=purchase status=200</pre>										Date tin	Q			
✓ ✓ 331 events (4/14/16 12:01:28.000 AM to 4/14/16 1:01:28.000 AM) No Event Sampling ✓ Job ✓ II ■ A 🖶 🛓												✓ Fast Mode ✓		
Events (331) Pati	terns	Statistic	s Visualizat	ion										
Format Timeline 🗸 🛛 –	Timeline - Zoom Out + Zoom to Selection × Deselect								1 minute per column					
				والموالي المراج										
		Lis	st 🗸 🖌 🖌 Format	✓ 20 Per Page ✓	< Prev	1 2	3	4	5 6	7 8	9	Next >		
< Hide Fields :	■ All Fields	i	Time	Event										
Selected Fields a host 3 a source 1		>	4/14/16 1:01:27.537 AM	175.135.152.171 - [13/Apr/2016 t_id=BW-3&JSESSIONID=SD7SBL4FF6AI _id=BW-3" "mozilla/5.0 (iPhone; U (KHTML, like Gecko) Version/5.0.2 host = webserver-03 source = /opt/apac	DFF3 HTTP 1.1" 200 998 U; CPU iPhone OS 4_2_6 2 Mobile/8E200 Safari/	8 "http: 5 like M 6533.18	//sho lac OS 8.5" 1	p.acmo X; ei 014	e.com/ n-us)	′produc AppleWe	t.screen?	product		
a sourcetype 1 Interesting Fields a action 1		>	4/14/16 1:01:21.026 AM	191.97.163.32 - [13/Apr/2016 1] _id=MCF-3&JSESSIONID=SD10CSL3FF1 3" "Opera/9.80 (Android; Opera M. host = webserver-03 source = /opt/apac	OADFF7 HTTP 1.1" 200 4 ini/6.5.27452/26.1235;	89 "htt U; en)	p://s Pres	hop.a to/2.	cme.co 8.119	om/oldl: Version	ink?item_	id=MCF-		
 a address 100+ a customer_segment 7 a customer_value 3 a customer_wallet 3 		>	4/14/16 1:01:14.242 AM	71.127.156.144 [13/Apr/2016 3d778c4b3f9a&category=Misc&JSESS do?action=purchase&item_id=WPSS- ike Mac OS X; ja_JP) AppleWebKit 0:EBDV/iPhone2 1:EBMD/iPhone:EBS	IONID=SD4SCL2FF6ADFF6 2&product_id=WPSS-2" " (KHTML, like Gecko) M	HTTP 1. mozilla Nobile [1" 20 1/5.0 FBAN/	0 781 (iPho FBFor	"http ne; U; IPhone	o://shop CPU i ;FBAV/4	o.acme.co Phone OS 4.0.3;FBB	m/cart. 4_3_3 l V/4030.		

2: Plot Purchase Value as Metric Timeseries



3: Investigate Database Errors

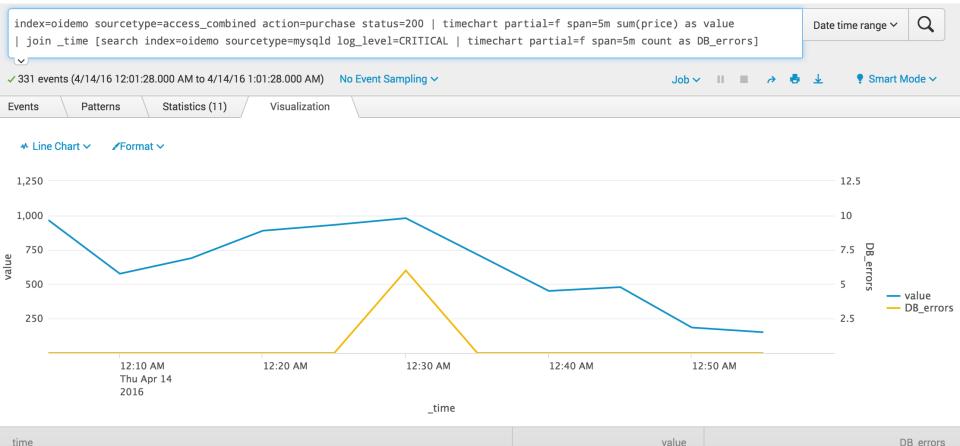
<pre>index=oidemo sourcetype=mysqld log_level=CRITICAL</pre>		Date time range 🗸		Q				
 ✓ ✓ 6 events (4/14/16 12:01:28.000 AM to 4/14/16 1:01:28.000 AM) No Event Sampling ✓ ✓ 6 events (4/14/16 12:01:28.000 AM to 4/14/16 1:01:28.000 AM) 								lode 🗸
Events (6) Patterns Statistics Visualization								
Format Timeline - Zoom Out + Zoom to Selection × Deselect							1 minute pe	r column

< Hide Fields	:≡ All Fields	i	Time	Event
Selected Fields		>	4/14/16 12:30:40.841 AM	<pre>13-Apr-2016 17:30:40:841177 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlo g/localhost-3306-bin.000020' (Errcode: 28). Waiting for someone to free space Retry in 60 secs host = mysql-02 { source = /usr/local/mysql/logs/mysqld.log { sourcetype = mysqld</pre>
a source 1 a sourcetype 1		>	4/14/16 12:30:40.706 AM	<pre>13-Apr-2016 17:30:40:706936 [CRITICAL] Error writing file '/mysqllog/slow_log/localhost_3306_slow_q ueries.log' (errno: 1) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysqld</pre>
Interesting Fields # date_hour 1 # date_mday 1		>	4/14/16 12:30:40.695 AM	<pre>13-Apr-2016 17:30:40:695540 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlo g/localhost-3306-bin.000020' (Errcode: 28). Waiting for someone to free space Retry in 60 secs host = mysql-02 { source = /usr/local/mysql/logs/mysqld.log { sourcetype = mysqld</pre>
<pre># date_minute 1 a date_month 1 # date_second 1</pre>		>	4/14/16 12:30:40.489 AM	<pre>13-Apr-2016 17:30:40:489892 [CRITICAL] Error writing file '/mysqllog/binlog/localhost-3306-bin' (er rno: 28) host = mysql-02 { source = /usr/local/mysql/logs/mysqld.log { sourcetype = mysqld</pre>

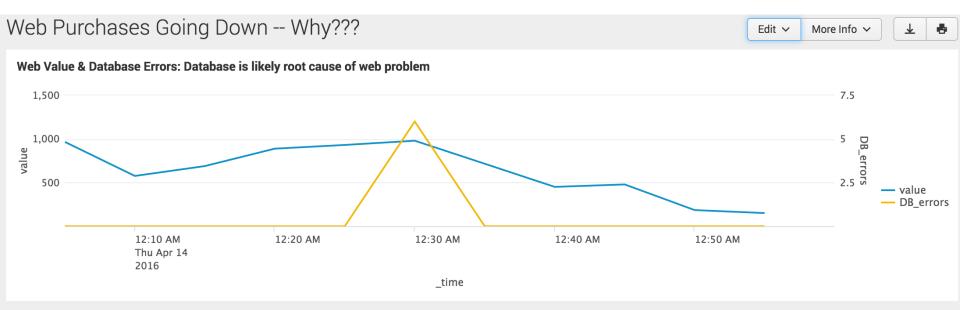
4: Plot Database Errors

	ndex=oidemo sourcetype=mysqld log_le	Date time range 🗸	Q				
✓ 6	5 events (4/14/16 12:01:28.000 AM to 4/14/16 1:	01:28.000 AM) No Ever	nt Sampling 🗸	J	Job 🗸 🔢 🔳 🤌 🖶	🖳 🛓 🍷 Smart	Mode 🗸
Eve	ents Patterns Statistics (11)	Visualization					
DB_errors	▲ Line Chart ∨	12:20 AM	12:30 AM	12:40 AM	12:50 AM		• DB_errors
	2016		_time				
_tir	ne						DB_errors

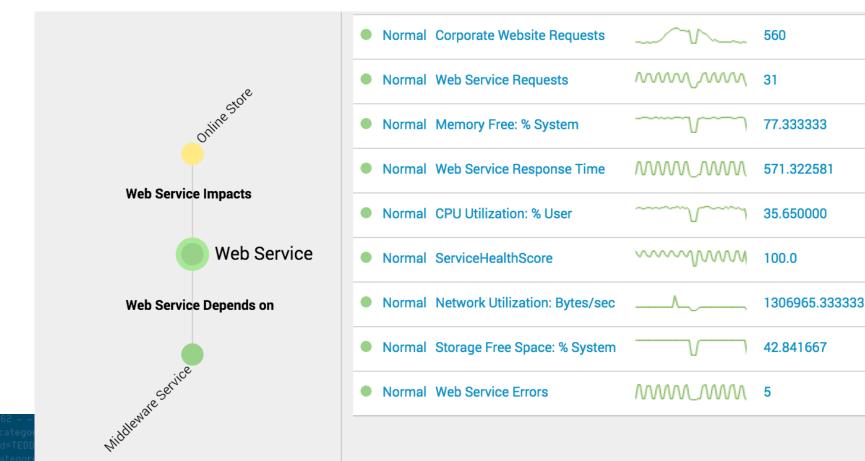
5: Correlate DB problems with purchase value



6: Save as Deliverable, Send to Stakeholders



7: Move toward proactive monitoring stance



Copyright © 2016 Splunk Inc.

Putting It All Together: **Doing Data Science**

-0.251.81 - [02/Feb/2011e Screen?category_ide/E0004W20

1.501 (Contractory) (Contracto tategory_id=FLOWERS: Mozile40 (or SEASES OF SEASES AND A COMPANY OF THE ADDRESS OF SEASES ategory_id=TEDDY* +10200af400 (competition id=GIFTS8.JSESSIONO=S015L6F5AUF801 egory_id=GIFTS+ Mozilla/5.0 (Window; 1, Wei SONID=SD4SL5FF2ADFF1 HTTP 11*40 200 STATE RP-LI-02* Googlebot/2.1 (http://www.goodeen reen?category_id=FLOWERS&RESAME-Signa zilla/5.0 (Macintosh; U; Intel Nac 05 % 10.6.3 extinguisting) GET /product.screen?product_id=1-Sit-464.game.equ temld=EST-16&product_id=FI-5W-dt+Weak50 (Koldet Nation 228.251.81 - - [02/Feb/2011.16:0023] & Ipoint-machinesso

y.screen?category_id=TEDOY' NozilaSU (Namet, Namet, Namet, Namet, Namet, Namet, Namet, Namet, Namet, Namet, Nam 62 - - [02/Feb/2011:16:00:23] @T /productsreadpoint (Reserv category_id=FLOWERS* Mozilla/4.0 (compilite NEGE Wolds: %) d=TEDDY8:JSESSIONID=SD9SL4FF4ADFF8 KTP 12:00 XXX MARE

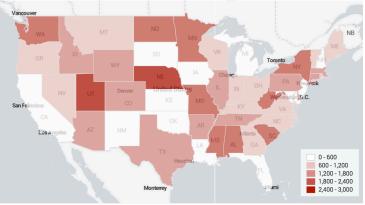
ategory_id=TEDDY" Mozilla/4.0 (compatible; NEE 6); Takin: BPA id=GIFTS&JSESSIONID=SDISL&FF3AUFR ITTI 12 # 28 1444 egory_id=GIFTS* Nozilla/5.0 (Windows; ال Windows) المعالي المعالي المعالي المعالي المعالي المعالي المعالي المعا SIONID=SD4SL5FF2ADFF1 HTTP 11 400 29R @Taol RP-LI-02* Googlebot/21 (http://mm.gov reen?category_id=FLOWERS&&ESSION-9884 rilla/5.0 (Macintosh; U; Intel Nac (6 x 10,63 and GET /product.screen?product.id=Fl.sH.084 temld=EST-16&product_id=FFSW-UP https://www. 228.251.81 - - [02/Feb/201116:00:3] & 1.screen?category_id=TED0% Mosks60 52 - - [02/Feb/2011.16.0023] @1/p ategory_id=FLOWERS' Mozilla/40 H=TEDDY&JSESSIONID=SD9SL4F400F

tegory_id=TEODY' Nozilla/4.0 (con CUETSS, ISESSIONDESDISLOFFSA



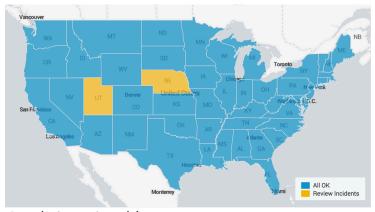
Emphasize <u>Actionable</u> Insights

- Avoid eye-candy visualizations
 - "Laser beam" threat dashboards look cool but are useless
- "How does this help me solve my problem?"
- Guide the viewer to drilldown & act quickly



Confusing viz: not actionable

S2 - - [02/Feb/2011:16:00:23] 'GET /product_screen?product_iHFFFW (20:3300 90/516:00) With/image



Good viz: actionable



Doing Data Science

- Data Scientists resist easy characterization. A bit of:
 - Statistician
 - Software Engineer
 - Business Analyst
 - Spock on the bridge
- Many scales of action:
 - Get your hands dirty when needed
 - But step back and see the big picture
 - Emphasize "actionable insights" throughout the whole org
 - Also have political credibility to say, "This is a bad decision, don't do it"
- Data Scientists guide organizations through building data products to solve big problems and deliver value



