

A Mathematical Perspective on Data Science

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

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

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 data-savvy 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 \rightarrow Data
 - Example problem solving
- Putting It All Together: Doing Data Science
 - Emphasize actionable insights
 - Tie it together to deliver “value” to stakeholders

- # Agenda
- Basic Definitions:
 - Data, Stakeholders, Use Cases, Deliverables
 - Query languages as functional programming
 - Every query is a map $f: \text{Data} \rightarrow \text{Data}$
 - Example problem solving
 - Putting It All Together: Doing Data Science
 - Emphasize actionable insights
 - Tie it together to deliver “value” to stakeholders

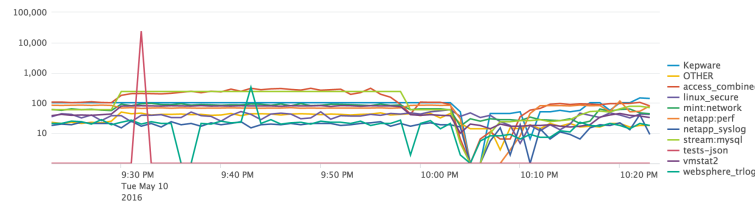


Basic Definitions


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.)

```
199.70.25.81 - [10/May/2016 22:18:23:078770] "GET /oldlink?item_id=MCF-38J
SESSIONID=SD45CL2FF6ADFF2 HTTP/1.1" 404 948 "http://shop.acme.com/cart.do?ac
tion=remove&item_id=MCF-38product_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
- 
- Three green silhouettes of people, representing stakeholders, arranged in a group. The silhouettes are simple, rounded shapes with a white outline for the neck and head. They are positioned in the lower right area of the slide.

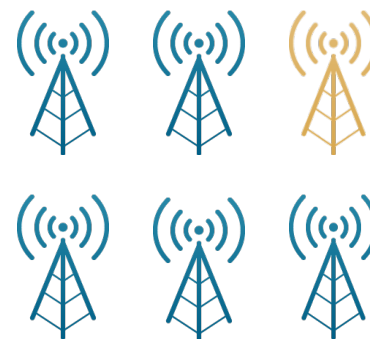


What is a Stakeholder? (cont.)

- 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 $\text{parent}(i) = A$)
 - "Value" modeled by objective functions ($U_{i,n}$ = objective # n for person i)
 - Is the action "pivotal" for the outcome? (ie $\mathbb{E}[U | \text{do}(a)] > \mathbb{E}[U | \text{do}(\text{not } a)]$?)
- 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)

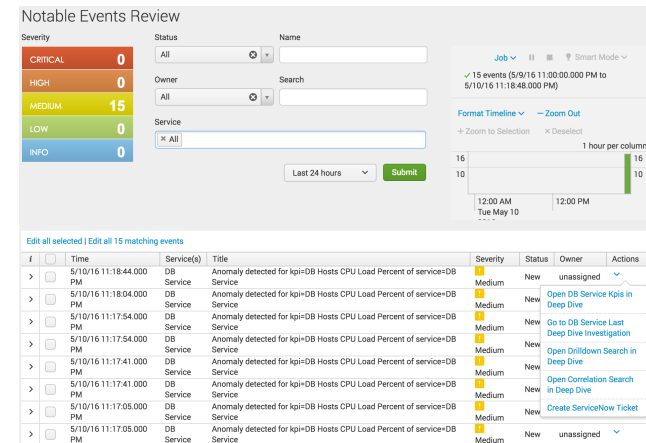
What is a Use Case?

- 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 \approx
 $\text{num outages} * \text{time to resolution} * \text{cost of labor}$
 - Stakeholders include IT & Business, and impacted customers



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





Queries as Functional Programming

Query Languages

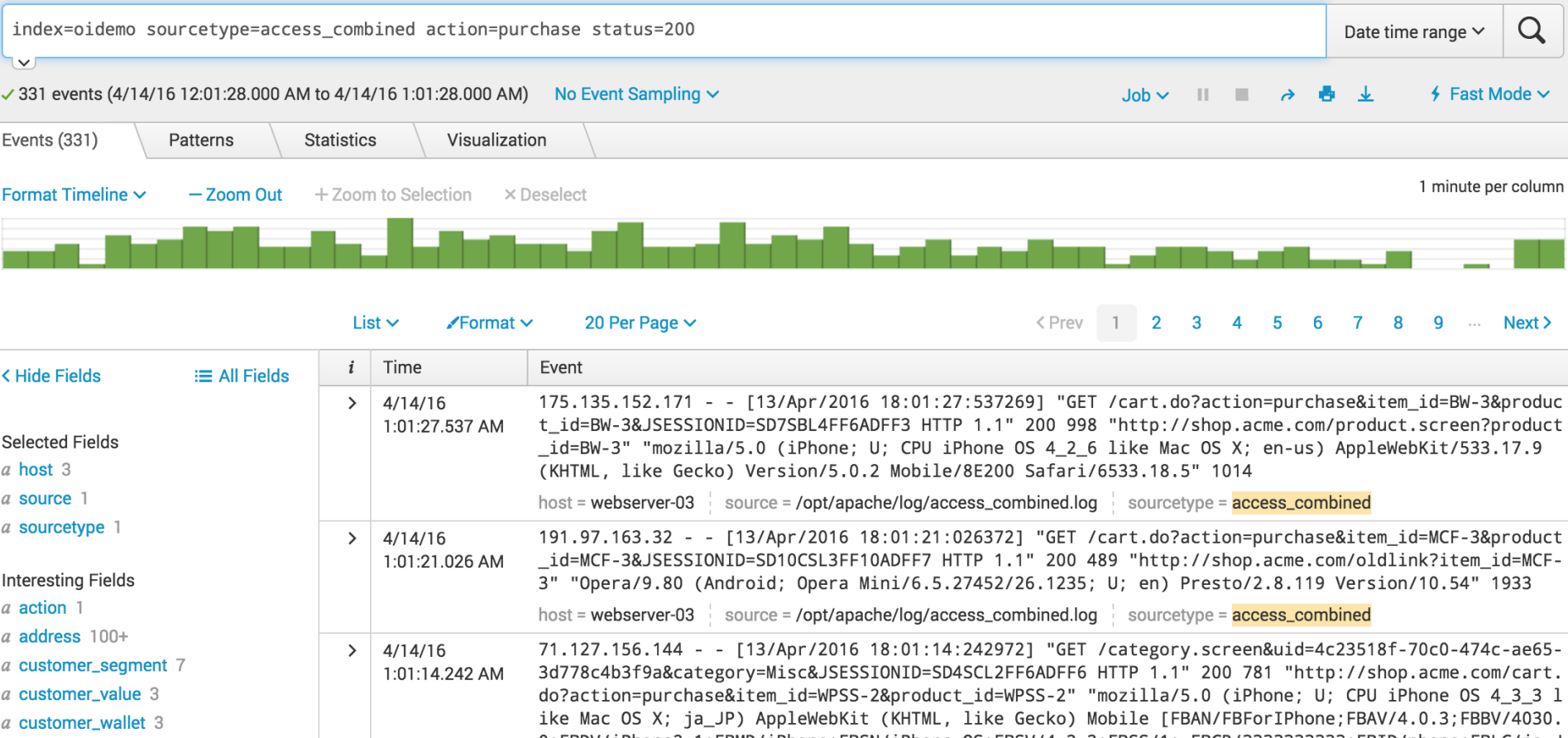
- Query languages provide a formulaic approach to working with data
- A "query" is a string which tells where to get the data, what to do with the data, and where to put the data (incl visualization or DB)
- Mathematically, every query is a FUNCTION $f : \text{Data} \rightarrow \text{Data}$
- Queries can be composed (with | symbol), and analysis is iterative
- Example 1: Successful Purchase Actions from Web Logs

```
sourcetype=access combined action=purchase status=200
```

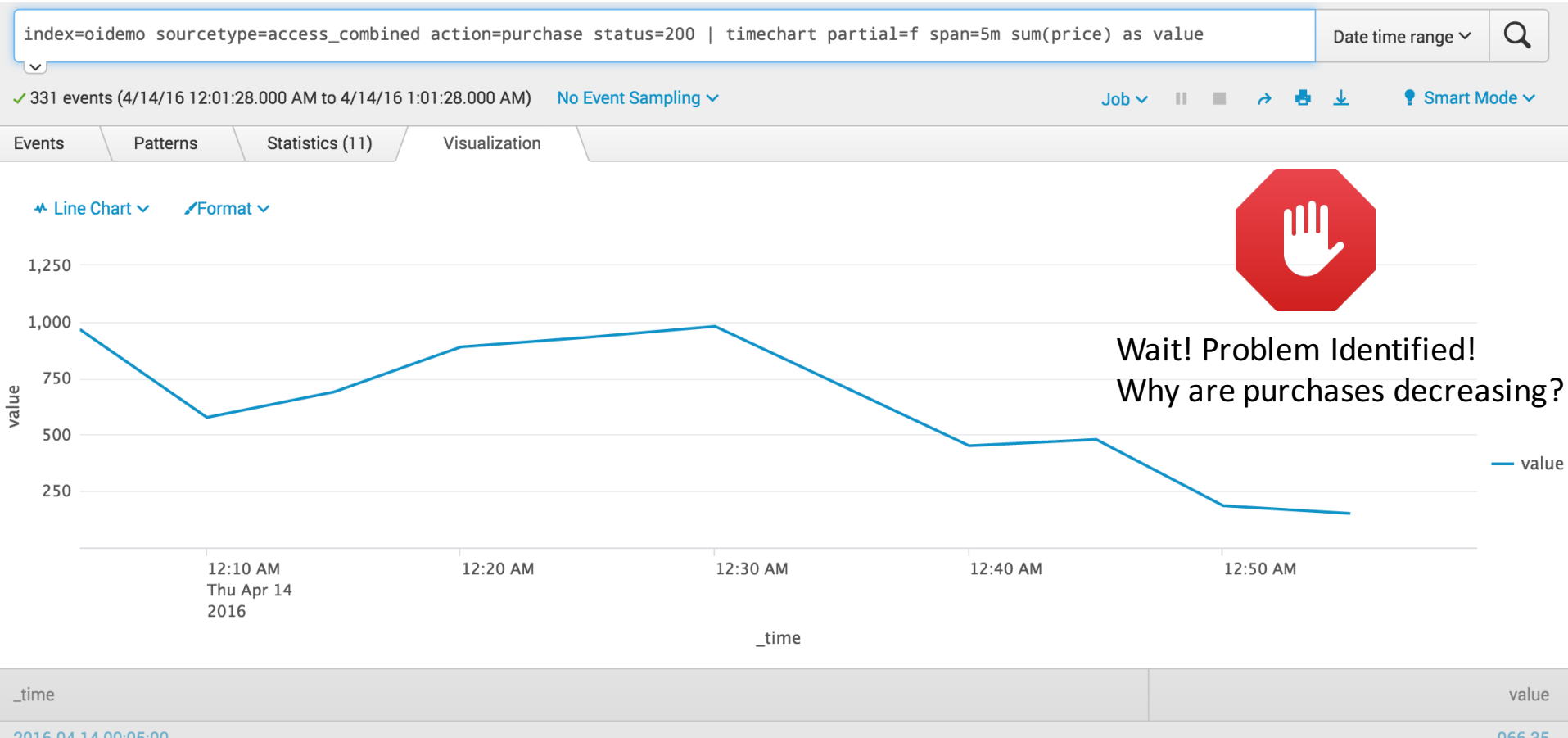
- Example 2: Plot Purchase Value as Metric Timeseries

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

1: Successful Purchase Actions from Web Logs



2: Plot Purchase Value as Metric Timeseries



3: Investigate Database Errors

Date time range ▾ 🔍

✓ 6 events (4/14/16 12:01:28.000 AM to 4/14/16 1:01:28.000 AM) No Event Sampling ▾

Job ▾ ⏸ ■ ➡ 🖨 ⬇ ⚡ Smart Mode ▾

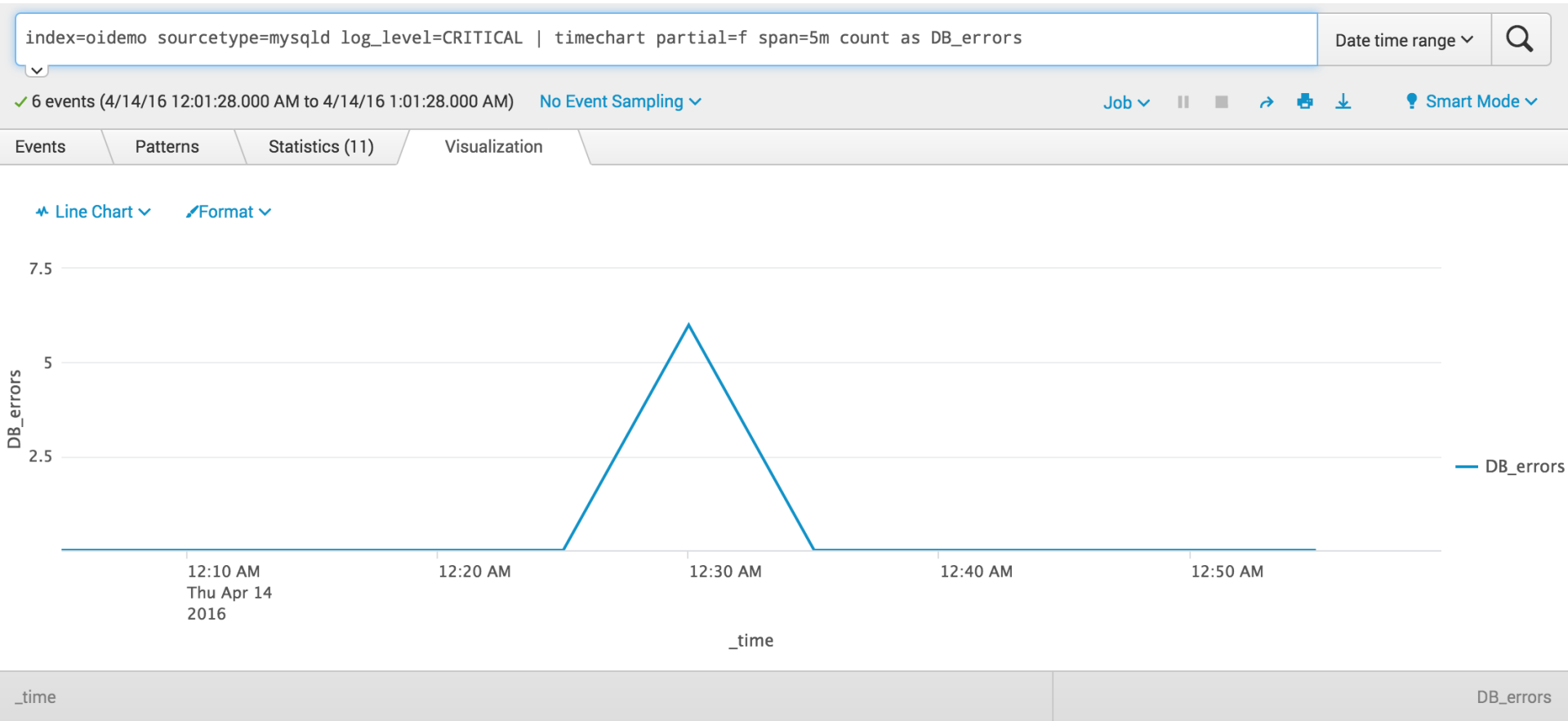
Events (6) Patterns Statistics Visualization

Format Timeline ▾ — Zoom Out + Zoom to Selection × Deselect 1 minute per column

List ▾ ↗ Format ▾ 20 Per Page ▾

<div>< Hide Fields</div> <div>Selected Fields</div> <div>a host 1</div> <div>a source 1</div> <div>a sourcetype 1</div> <div>Interesting Fields</div> <div># date_hour 1</div> <div># date_mday 1</div> <div># date_minute 1</div> <div>a date_month 1</div> <div># date_second 1</div>	<div>≡ All Fields</div>	<table><thead><tr><th>i</th><th>Time</th><th>Event</th></tr></thead><tbody><tr><td>></td><td>4/14/16 12:30:40.841 AM</td><td>13-Apr-2016 17:30:40:841177 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql</td></tr><tr><td>></td><td>4/14/16 12:30:40.706 AM</td><td>13-Apr-2016 17:30:40:706936 [CRITICAL] Error writing file '/mysqllog/slow_log/localhost_3306_slow_queries.log' (errno: 1) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql</td></tr><tr><td>></td><td>4/14/16 12:30:40.695 AM</td><td>13-Apr-2016 17:30:40:695540 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql</td></tr><tr><td>></td><td>4/14/16 12:30:40.489 AM</td><td>13-Apr-2016 17:30:40:489892 [CRITICAL] Error writing file '/mysqllog/binlog/localhost-3306-bin' (errno: 28) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql</td></tr></tbody></table>	i	Time	Event	>	4/14/16 12:30:40.841 AM	13-Apr-2016 17:30:40:841177 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql	>	4/14/16 12:30:40.706 AM	13-Apr-2016 17:30:40:706936 [CRITICAL] Error writing file '/mysqllog/slow_log/localhost_3306_slow_queries.log' (errno: 1) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql	>	4/14/16 12:30:40.695 AM	13-Apr-2016 17:30:40:695540 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql	>	4/14/16 12:30:40.489 AM	13-Apr-2016 17:30:40:489892 [CRITICAL] Error writing file '/mysqllog/binlog/localhost-3306-bin' (errno: 28) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql
i	Time	Event															
>	4/14/16 12:30:40.841 AM	13-Apr-2016 17:30:40:841177 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql															
>	4/14/16 12:30:40.706 AM	13-Apr-2016 17:30:40:706936 [CRITICAL] Error writing file '/mysqllog/slow_log/localhost_3306_slow_queries.log' (errno: 1) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql															
>	4/14/16 12:30:40.695 AM	13-Apr-2016 17:30:40:695540 [CRITICAL] /opt/mysql/bin/mysqld: Disk is full writing '/mysqllog/binlog/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 = mysql															
>	4/14/16 12:30:40.489 AM	13-Apr-2016 17:30:40:489892 [CRITICAL] Error writing file '/mysqllog/binlog/localhost-3306-bin' (errno: 28) host = mysql-02 source = /usr/local/mysql/logs/mysqld.log sourcetype = mysql															

4: Plot Database Errors



5: Correlate DB problems with purchase value

```
index=oidemo sourcetype=access_combined action=purchase status=200 | timechart partial=f span=5m sum(price) as value  
| join _time [search index=oidemo sourcetype=mysqlld log_level=CRITICAL | timechart partial=f span=5m count as DB_errors]
```

Date time range ▾



✓ 331 events (4/14/16 12:01:28.000 AM to 4/14/16 1:01:28.000 AM)

No Event Sampling ▾

Job ▾



Smart Mode ▾

Events

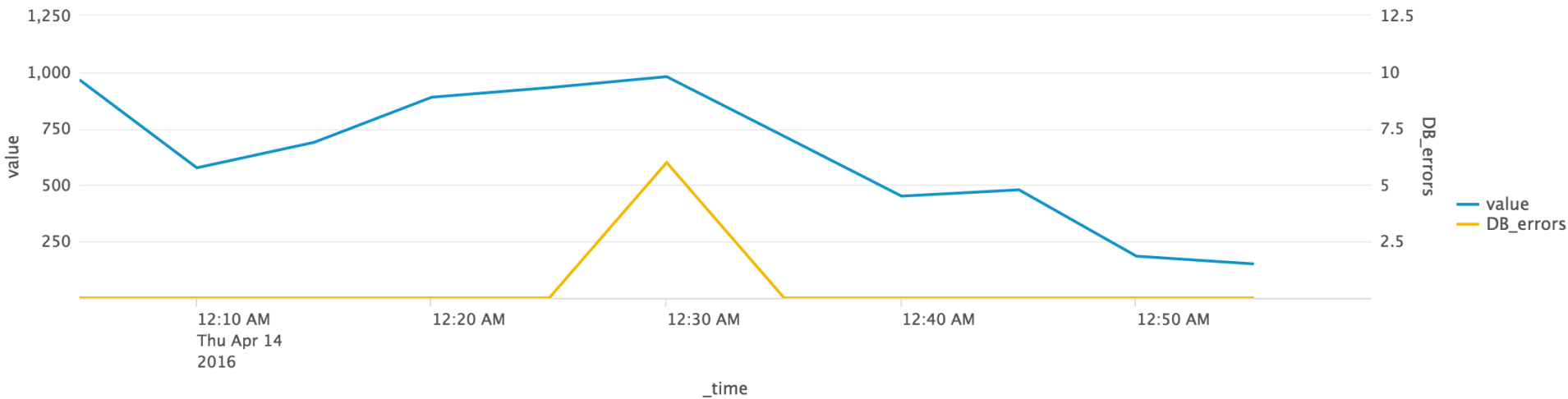
Patterns

Statistics (11)

Visualization

Line Chart ▾

Format ▾



_time

value

DB_errors

6: Save as Deliverable, Send to Stakeholders

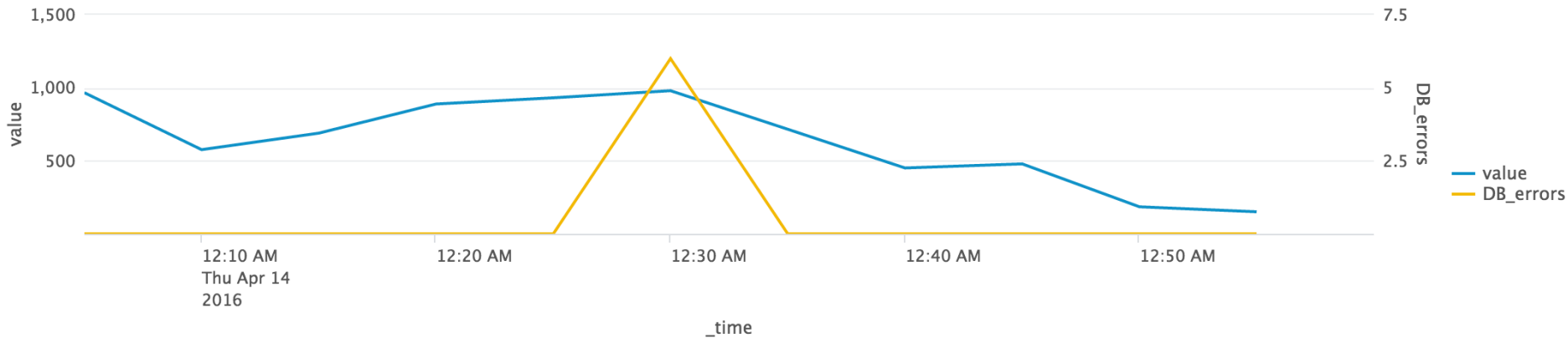
Web Purchases Going Down -- Why???

Edit ▾

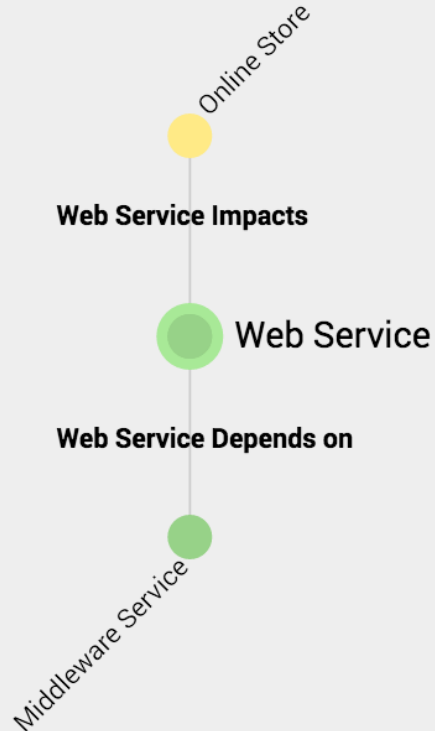
More Info ▾












Web Value & Database Errors: Database is likely root cause of web problem



7: Move toward proactive monitoring stance



● Normal	Corporate Website Requests		560
● Normal	Web Service Requests		31
● Normal	Memory Free: % System		77.333333
● Normal	Web Service Response Time		571.322581
● Normal	CPU Utilization: % User		35.650000
● Normal	ServiceHealthScore		100.0
● Normal	Network Utilization: Bytes/sec		1306965.333333
● Normal	Storage Free Space: % System		42.841667
● Normal	Web Service Errors		5



Putting It All Together: Doing Data Science

[illegible]

- ```
82 - [02/Feb/2011:16:00:23] "GET /product.screen?product_id=FI-FW-429L&SESSIONID=SD9SL4FF4ADFF8 HTTP/1.1" 200 2100 http://www.myflowershop.com/category/screen/category_screen.html?category_id=FLOWERS* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896) Safari/525.4 197.107.2.114
category_id=FLOWERS* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896) http://www.myflowershop.com/category/screen/category_screen.html?category_id=TEDDY&SESSIONID=SD9SL4FF4ADFF8 HTTP/1.1" 200 3439 Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896) http://www.myflowershop.com/category/screen/category_screen.html?category_id=TEDDY* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896) http://www.myflowershop.com/category/screen/category_screen.html?category_id=TEDDY* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896) http://www.myflowershop.com/category/screen/category_screen.html?category_id=TEDDY* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322.3896)
```

[illegible][illegible]

62 - - [02/Feb/2011:16:00:23] "GET /productscreen?product\_id=FW-429LJSESSID=SD9SL4FF4ADFF8 HTTP 1.1" 200 3439 Windows NT 5.1; SV1; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4332.1034) http://www.myflowershop.com/categoryscreen?category\_id=FLOWERS\*  
category\_id=FLOWERS\* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4332.1034) http://www.myflowershop.com/categoryscreen?category\_id=FLOWERS\*  
d=TEDDY&JSESSIONID=SD9SL4FF4ADFF8 HTTP 1.1" 200 3439 Windows NT 5.1; SV1; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4332.1034) http://www.myflowershop.com/categoryscreen?category\_id=TEDDY\*  
category\_id=TEDDY\* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4332.1034) http://www.myflowershop.com/categoryscreen?category\_id=TEDDY\*

- 
- A cartoon illustration of two men. The man on the left, wearing a green shirt, is shouting into a large orange megaphone. The man on the right, wearing a blue shirt, is holding his hand to his ear, looking surprised or concerned.