Understanding the process of data analysis by studying interaction records

Sara Alspaugh
@salspaugh
AMPLab, EECS Department, UC Berkeley

in collaboration with:
Archana Ganapathi (Splunk),
Beidi Chen, Jessica Lin, Marti Hearst, Randy Katz (UC Berkeley)
personalization
system improvement
interface optimization
business intelligence
usage characterization

guidelines for data exploration

Perer and Shneiderman. “Systematic yet flexible discovery: guiding domain experts through exploratory data analysis.”

intelligent help systems

Horvitz et al. “The Lumière project: Bayesian user modeling for inferring the goals and needs of software users.”

predictive interfaces

Kandel et al. “Wrangler: interactive visual specification of data transformation scripts.”

automatic analysis generation

Jock Mackinlay. “Automating the Design of Graphical Presentations of Relational Information.”

Fig. 10. The Encodes relationship in the RML language. The graphical sentence is on the right. The gray lines indicate that the middle term is encoded by the objects, and the tuples are encoded by the relative positions of the marks.
Outline

• Background

• Data collected

• Challenges

• Lessons learned
Outline

• Background

• Data collected

• Challenges

• Lessons learned
Data analysis in the abstract:

Example input:


Example operations:

```
SELECT COUNT(*) FROM table;
```

```
extract_event_fields(s)
```

```
add_series(data)
```

```
render_graph(g)
```

Example output:

```
(5738,)
```

```
[127.0.0.1, frank, 200]
```
Many operations are provided.

input data → operation → output data
An example set of operations:

input data → filter → output data

aggregate

join

visualize
State machine view

- filter
- aggregate
- join
- visualize
Data flow view
What are the most common paths through this graph?
Given one action, what action is most likely to be taken next?
What additional information should we examine?
What are the most common paths through this graph?

Given one action, what action is most likely to be taken next?

What are the challenges in examining these questions?

What additional information should we examine?
Outline

• Background

• Data collected

• Challenges

• Lessons learned
for example:

business analyst
op engineer
data scientist
actual scientist

for example:

Excel
Tableau
R
Splunk
Analyzing Log Analysis

• Splunk is a platform for analysis and visualization of semi-structured log data
• collected >200K queries to analyze
• queries have CL utilities-like syntax

## Analyzing Log Analysis

### Splunk Search Query

```
source=*access_log* | eval http_success = if(status=200, "true", "false") | timechart count by http_success
```

### Results

- **Events:** 618,369 events (before 3/31/14 10:57:28.000 PM)
- **Format:** Timeline, Zoom Out, Zoom to Selection, Deselect

### Selected Fields

- **Host:** 1
- **Source:** 1
- **Source Type:** access_combined

### Interesting Fields

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host</td>
<td>ronnie.sv.splunk.com</td>
</tr>
<tr>
<td>Source</td>
<td>/home/aganapath/access.log.1.gz</td>
</tr>
<tr>
<td>Source Type</td>
<td>access_combined</td>
</tr>
</tbody>
</table>

### Event Details

1. **12/20/09 4:02:12.000 AM**
   - **Event:** `GET /base/Documentation HTTP/1.1 200 20551 "Mozilla/5.0"`
   - **Details:** `host = ronnie.sv.splunk.com | source = /home/aganapath/access_log.1.gz | source_type = access_combined`

2. **12/20/09 4:02:12.000 AM**
   - **Event:** `GET /base/Documentation/3.4.11/Admin/Regmonfiltersconf///index.php?SERVER[DOCUMENT_ROOT]=http://www.dconroll.co.kr/data/members/id1.txt HTTP/1.1 302 "Mozilla/5.0"`
   - **Details:** `host = ronnie.sv.splunk.com | source = /home/aganapath/access_log.1.gz | source_type = access_combined`

3. **12/20/09 4:02:11.000 AM**
   - **Event:** `GET /web_assets/v4/css/styles.css HTTP/1.1 200 730 "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/20000101 Firefox/3.0.9 Win32; IEEE 1394 v4"`
   - **Details:** `host = ronnie.sv.splunk.com | source = /home/aganapath/access_log.1.gz | source_type = access_combined`

4. **12/20/09 4:02:11.000 AM**
   - **Event:** `GET /themes/splunk_conscripts/js/slimbox2.js?12-03-2009 HTTP/1.1 200 1988 "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/20000101 Firefox/3.0.9 Win32; IEEE 1394 v4"`
   - **Details:** `host = ronnie.sv.splunk.com | source = /home/aganapath/access_log.1.gz | source_type = access_combined`

5. **12/20/09 4:02:09.000 AM**
   - **Event:** `GET /themes/splunk_conscripts/js/global.js?12-03-2009 HTTP/1.1 200 5106 "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/20000101 Firefox/3.0.9 Win32; IEEE 1394 v4"`
   - **Details:** `host = ronnie.sv.splunk.com | source = /home/aganapath/access_log.1.gz | source_type = access_combined`
Analyzing Log Analysis

search "error"
| stats count by status
| lookup statuscodes status OUTPUT statusdesc

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
<th>statusdesc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404</td>
<td>Not Found</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>Internal Server Error</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
</tr>
</tbody>
</table>
Exploring Data Exploration

• Tableau is desktop tool for visual analysis and presentation of relational data
• collected 25 cases: logs plus transcripts
• use cases contain activity and motivations
Exploring Data Exploration
Exploring Data Exploration

{
  "action": "tabui:drop-ui",
  "parameters": {
    "drag-description": "",
    "drag-source": "drag-drop-schema",
    "drop-target": "drag-drop-viz",
    "field-encodings": [
      {
        "encoding-type": "invalid-encoding",
        "fn": "[On_Time_On_Time_Performance_2001_9 Extract].[sum:Number of Records:qk]"
      }
    ],
    "is-copy": "false",
    "is-dead-drop": "false",
    "is-right-drag": "false",
    "shelf-drag-source-position": {
      "is-override": false
    },
    "shelf-drop-context": "none",
    "shelf-drop-target-position": {
      "is-overlay": true,
      "is-override": false,
      "shelf-pos-index": 0,
      "shelf-type": "rows-shelf"
    },
    "target-sheet": "Sheet 1"
  },
  "timestamp": 1409793051.437
}
Outline

• Background

• Data collected

• Challenges

• Lessons learned
Challenges for analysis

• very large transformation space
• larger space of sequences of transformations
• mismatch between representation of transformations and analysis questions
• missing context information
  – data, metadata, system state, etc.
• complex structure of heterogeneous activities - challenging to analyze, featurize, and model
Challenges for analysis

• very large transformation space
• larger space of sequences of transformations
• mismatch between representation of transformations and analysis questions
• missing context information
  – data, metadata, system state, etc.
• complex structure of heterogeneous activities - challenging to analyze, featurize, and model
<table>
<thead>
<tr>
<th>command</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>search</td>
<td>232373</td>
</tr>
<tr>
<td>eval</td>
<td>178080</td>
</tr>
<tr>
<td>stats</td>
<td>75927</td>
</tr>
<tr>
<td>table</td>
<td>44967</td>
</tr>
<tr>
<td>fields</td>
<td>37803</td>
</tr>
<tr>
<td>rename</td>
<td>35919</td>
</tr>
<tr>
<td>where</td>
<td>32402</td>
</tr>
<tr>
<td>inputlookup</td>
<td>30490</td>
</tr>
<tr>
<td>sort</td>
<td>30442</td>
</tr>
<tr>
<td>lookup</td>
<td>28620</td>
</tr>
<tr>
<td>outputlookup</td>
<td>27042</td>
</tr>
<tr>
<td>dedup</td>
<td>22731</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>localop</td>
<td>27</td>
</tr>
<tr>
<td>reverse</td>
<td>15</td>
</tr>
<tr>
<td>abstract</td>
<td>10</td>
</tr>
<tr>
<td>map</td>
<td>7</td>
</tr>
<tr>
<td>anomalies</td>
<td>3</td>
</tr>
<tr>
<td>extract</td>
<td>2</td>
</tr>
<tr>
<td>outlier</td>
<td>2</td>
</tr>
<tr>
<td>datamodel</td>
<td>2</td>
</tr>
<tr>
<td>format</td>
<td>1</td>
</tr>
<tr>
<td>outputtext</td>
<td>1</td>
</tr>
<tr>
<td>dbinspect</td>
<td>1</td>
</tr>
<tr>
<td>action</td>
<td>count</td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>tabui:drop-ui</td>
<td>506</td>
</tr>
<tr>
<td>tabdoc:undo</td>
<td>165</td>
</tr>
<tr>
<td>tabdoc:show-me</td>
<td>141</td>
</tr>
<tr>
<td>tabdoc:move-dashboard-edge</td>
<td>90</td>
</tr>
<tr>
<td>tabui:change-aggregation-ui</td>
<td>75</td>
</tr>
<tr>
<td>tabdoc:highlight</td>
<td>72</td>
</tr>
<tr>
<td>tabui:save-workbook</td>
<td>72</td>
</tr>
<tr>
<td>tabdoc:rename-sheet</td>
<td>68</td>
</tr>
<tr>
<td>tabdoc:select-legend-item</td>
<td>55</td>
</tr>
<tr>
<td>tabdoc:quantitative-quick-filter-edit</td>
<td>53</td>
</tr>
<tr>
<td>tabdoc:sort-from-indicator</td>
<td>53</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>tabui:view-title-card</td>
<td>1</td>
</tr>
<tr>
<td>tabui:publish-workbook-to-workgroup</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:remove-field-label</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:on-toggle-axis-ranges</td>
<td>1</td>
</tr>
<tr>
<td>tabui:import-workbook-ui</td>
<td>1</td>
</tr>
<tr>
<td>tabui:save-public</td>
<td>1</td>
</tr>
<tr>
<td>tabui:select-parent-zone</td>
<td>1</td>
</tr>
<tr>
<td>tabui:create-set-ui</td>
<td>1</td>
</tr>
<tr>
<td>tabui:add-reference-line-ui</td>
<td>1</td>
</tr>
<tr>
<td>tabui:export-packaged</td>
<td>1</td>
</tr>
<tr>
<td>tabui:reset-cards</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:group-by-table</td>
<td>1</td>
</tr>
<tr>
<td>action</td>
<td>popularity</td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>tabui:drop-ui</td>
<td>11</td>
</tr>
<tr>
<td>tabui:connection-analyze-data-ui</td>
<td>10</td>
</tr>
<tr>
<td>tabui:document</td>
<td>10</td>
</tr>
<tr>
<td>tabdoc:show-me</td>
<td>10</td>
</tr>
<tr>
<td>tabui:new-worksheet-ui</td>
<td>10</td>
</tr>
<tr>
<td>tabui:change-aggregation-ui</td>
<td>9</td>
</tr>
<tr>
<td>tabdoc:undo</td>
<td>9</td>
</tr>
<tr>
<td>tabui:upgrade-datasources-ui</td>
<td>9</td>
</tr>
<tr>
<td>tabui:connect-datasource</td>
<td>9</td>
</tr>
<tr>
<td>tabdoc:select-legend-item</td>
<td>9</td>
</tr>
<tr>
<td>tabdoc:clear-legend-selection</td>
<td>8</td>
</tr>
<tr>
<td>tabui:new-dashboard-ui</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>tabui:edit-reference-line</td>
<td>1</td>
</tr>
<tr>
<td>tabui:trend-lines-ui</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:edit-story-point-caption</td>
<td>1</td>
</tr>
<tr>
<td>tabui:select-parent-zone</td>
<td>1</td>
</tr>
<tr>
<td>tabui:create-set-ui</td>
<td>1</td>
</tr>
<tr>
<td>tabui:export-packaged</td>
<td>1</td>
</tr>
<tr>
<td>tabui:view-legend</td>
<td>1</td>
</tr>
<tr>
<td>tabui:reset-cards</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:group-by-table</td>
<td>1</td>
</tr>
<tr>
<td>tabdoc:show-legend-on-dashboard</td>
<td>1</td>
</tr>
<tr>
<td>tabui:new-storyboard-ui</td>
<td>1</td>
</tr>
</tbody>
</table>
### Splunk top commands

<table>
<thead>
<tr>
<th>command</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>search</td>
<td>232373</td>
</tr>
<tr>
<td>eval</td>
<td>178080</td>
</tr>
<tr>
<td>stats</td>
<td>75927</td>
</tr>
<tr>
<td>table</td>
<td>44967</td>
</tr>
<tr>
<td>fields</td>
<td>37803</td>
</tr>
<tr>
<td>rename</td>
<td>35919</td>
</tr>
</tbody>
</table>

### Tableau top commands

<table>
<thead>
<tr>
<th>action</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>tabui:drop-ui</td>
<td>506</td>
</tr>
<tr>
<td>tabdoc:undo</td>
<td>165</td>
</tr>
<tr>
<td>tabdoc:show-me</td>
<td>141</td>
</tr>
<tr>
<td>tabdoc:move-dashboard-edge</td>
<td>90</td>
</tr>
<tr>
<td>tabui:change-aggregation-ui</td>
<td>75</td>
</tr>
</tbody>
</table>
Solution: hierarchical categorization of tasks, sub-tasks, lower-level activities

- Add info
- Append cols
- Bin
- Bucket
- Evaluate
- Extract
- IP location
- KV
- Output text
- Range map
- Regex
- Spath
- strcat
- XML KV

- Add col totals
- Count table
- Event count
- Geostats
- Stats
- Time chart
- Top

**Filter**

![Filter](image)

**Augment**

![Augment](image)

**Aggregate**

![Aggregate](image)
N = 82680 stages

N = 17085 queries
Splunk transformation Markov diagram
Challenges for analysis

- very large transformation space
- larger space of sequences of transformations
- mismatch between representation of transformations and analysis questions
- missing context information
  - data, metadata, system state, etc.
- complex structure of heterogeneous activities - challenging to analyze, featurize, and model
search "error"
| stats count by status
| lookup statuscodes status OUTPUT statusdesc

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
<th>statusdesc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404</td>
<td>Not Found</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>Internal Server Error</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
</tr>
</tbody>
</table>

search "error"
search "error"
| stats count by status
| lookup statuscodes status OUTPUT statusdesc

---

**Missing data**

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
<th>statusdesc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404</td>
<td>Not Found</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>Internal Server Error</td>
</tr>
</tbody>
</table>

---

0.0  - error 404
0.7  - error 500
<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Modifiers</th>
<th>Storage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>integer</td>
<td>not null default nextval('queries_id_seq'::regclass)</td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>text</td>
<td>text</td>
<td>not null</td>
<td>extended</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>double precision</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>is_interactive</td>
<td>boolean</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>is_suspicious</td>
<td>boolean</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>execution_time</td>
<td>double precision</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>earliest_event</td>
<td>double precision</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>latest_event</td>
<td>double precision</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>range</td>
<td>double precision</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>is_realtime</td>
<td>boolean</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>search_type</td>
<td>text</td>
<td></td>
<td>extended</td>
<td></td>
</tr>
<tr>
<td>splunk_search_id</td>
<td>text</td>
<td></td>
<td>extended</td>
<td></td>
</tr>
<tr>
<td>saved_search_name</td>
<td>text</td>
<td></td>
<td>extended</td>
<td></td>
</tr>
<tr>
<td>user_id</td>
<td>integer</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>session_id</td>
<td>integer</td>
<td></td>
<td>plain</td>
<td></td>
</tr>
</tbody>
</table>
source_type=solaris3-web-access

256 distinct queries
sourcetype=vmware:perf:*
Insufficient provenance

Repeat 1004 times

Queries presumed to be interactive
interarrival times

0.403
0.404
0.404
0.404
0.406
0.405
0.410
5.471
0.403
0.404
0.404
0.404
0.405
0.405
5.471
0.401
0.404
0.406
0.407
0.407
0.405
5.467

What’s wrong with this picture?
"action": "tabui:create-calculation-ui",
"timestamp": 1409793211.666,
"parameters": {
  "use-selector": "true"
}
## EDA Assignment 1 - First Foray

**Dataset:** Oakland City Budget

**Hypothesis:** The city spends most of its public safety budget (Police + Fire) on salaries.

### Exploration Log

<table>
<thead>
<tr>
<th>Goal / Question</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>See spending by department</td>
<td>Drag &quot;Amount&quot; to rows</td>
</tr>
<tr>
<td>What do all these fields mean?</td>
<td>Drag &quot;Department&quot; to Columns</td>
</tr>
<tr>
<td>Now I see &quot;Salaries&quot; in some values</td>
<td>Swap F/C</td>
</tr>
<tr>
<td>Now I see which are &quot;Expenses&quot; vs &quot;Revenues&quot;</td>
<td>New Sheet (2)</td>
</tr>
<tr>
<td>2. The biggest department is &quot;Non-Departmental&quot; - what's that about?</td>
<td>Drag &quot;Account Category&quot; to Rows</td>
</tr>
<tr>
<td></td>
<td>Add &quot;Account Type&quot;</td>
</tr>
<tr>
<td></td>
<td>Return to Sheet 1</td>
</tr>
<tr>
<td></td>
<td>Drag &quot;Account Type&quot; to Filters, include only &quot;Expense&quot;</td>
</tr>
</tbody>
</table>

Exploration transcripts for ground truth and missing information
Initial findings: EDA in Tableau

• Main challenges, opportunities for applying the results of our analysis:
  – Goal uncertainty: acting with an empty mental model
  – Complex ad hoc planned workflow management: keeping track of threads to investigate
  – Declaration-implementation mismatch: knowing what is wanted but not how to produce it
Challenges for analysis

- very large transformation space
- larger space of sequences of transformations
- mismatch between representation of transformations and analysis questions
- missing context information
  - data, metadata, system state, etc.
- complex structure of heterogeneous activities - challenging to analyze, featurize, and model
1. Featurize stages in given category.
2. Perform PCA.
3. Perform t-SNE.
Types of Filter Transformations

- Filters by long logical condition
- Filters by specifying string contains
- Filters by time range
- Deduplicates
- Uses macro
- Uses subsearch
- Uses regex
- More string-contains
- More selections, field-value searches
- Data set selections
- Sub-searches
- Field-value searches
- Simple contains-string searches
- Time range searches
- Function-based filters
- Macros
- Deduplication

Example searches:
- `search index=os`
- `search daysago=7`
- `search status=500 AND page=index.html`
- `search name=iexplore.exe OR name=Toolbar`
- `search name=process2 | fields tid`
- `search name=process*`
- `search alsapough`
- `where minutes/60<5`
Aggregation
- Visualize aggregation over time
- Aggregate, sort, limit
- Group by time

Types of Aggregate Transformations

- count, sort, limit
- top hostname limit="10"
- aggregate only
- stats sum(counts) AS sum
- aggregate events by time
- transaction startswith="--" endswith=" Time"
- aggregate, visualize only
- aggregate, visualize
- chart count by name
...work in this area has relied on either manually recorded provenance (e.g., user notes) or automatically recorded event-based insight provenance (e.g., clicks, drags, and key-presses), both approaches have fundamental limitations.”
Needed

• well-instrumented analysis tool with sizeable user base
• high-quality interaction records that contain context
• a third use case
Conclusion

• compelling applications of interaction records from data analysis tools
  – guidelines for data exploration
  – automatic analysis generation
  – intelligent help systems
  – predictive interfaces
  – ... and more!
• first step: exploring real-world analysis system interaction records
• barriers to analysis and modeling:
  – huge state space
  – complex and messy data sources
  – missing data
  – mismatch between data collection goal and analysis goal
• next step: building models for guidelines, prediction, and more
• other lessons:
  – need improved instrumentation and provenance
  – new techniques needed to better analyze complex, high-dimensional, and hierarchical activity data
  – possible benefits from co-designing analysis environment with environment model