Viska:
A System for Interactive Analysis of Performance Measurements

Helga Gudmundsdottir, Babak Salimi, Magda Balazinska, Dan Ports, and Dan Suciu

UW DB Day 2016
-- Work in Progress --
Performance Analysis

• **High level goal:** Want to **understand** and **analyze** performance characteristics of systems

• Systems **innovation, deployment** and **tuning** often involves **performance measurements**

• Need for **gathering** and **interpreting** large amounts of **system traces**
Motivating Scenarios

- **Comparing performance across systems**
  - Why is one system faster than another for this query?

- **Performance in distributed systems**
  - Why did my p90 latency go way up this week?

- **Performance tuning in database systems**
  - Why is this system configuration affecting the runtime?
Example

- **Task**: Investigate performance of new version of PostgreSQL
- **Workload**: Reading from a large fact table
Example

• **Task:** Investigate performance of new version of PostgreSQL

• **Workload:** Reading from a large fact table

• **Experiments:** Run simple queries on both versions in identical setups

  ```
  select [columns] from [table] where [...]
  ```

• **Vary:**
  - selected columns
  - size of table
  - filtering predicates
  - ...
Strawman Approach

OLD VERSION vs. NEW VERSION

Histogram of Old Version vs. New Version:
- Old version: PDF
  - Duration
  - Bars from 0 to 100
- New version: PDF
  - Duration
  - Bars from 0 to 100
Strawman Approach++

**Average Treatment Effect (ATE):**

\[ E(\text{duration}|\text{new version}) - E(\text{duration}|\text{old version}) \]

**Results:**

version \(\rightarrow\) duration
Strawman Approach++

**Average Treatment Effect (ATE):**

\[ E(\text{duration}|\text{new version}) - E(\text{duration}|\text{old version}) \]

**Results:** 14.10 seconds – 14.02 seconds = 0.08 seconds
Strawman Approach++

Average Treatment Effect

\[ E(\text{duration} \mid \text{new version}) - E(\text{duration} \mid \text{old version}) \]

Results: 14.10 seconds - 14.02 seconds = 0.08 seconds

Are these results valid?
Viska
(“wisdom” is Icelandic, “to see” in Lojban)

Principled approach to **generating**, **exploring** and **analyzing**
performance measurements
Experiment Design

- **Input variables**: The set of parameters we control and vary during experiments.

- **Observed variables**: All potentially relevant features that could influence our population.

- **Outcome variables**: Designated set of observed variables to be treated as outcome.

```python
# population
input_variables:
- system: postgres
- benchmark: tpcch-data
- scale_factor: [1,2,4,6,8,10]
- n_predicates: [0,1,2]
- in_columns: [10,17]
- clear_os_cache: [True, False]

outcome_variables:
- duration_s

observed_variables:
- in_rows
- out_rows
- in_columns
- ...
```
Exploring Results

```
from viska import Viska

v = Viska("parameters.yaml")
v.data.head()
```

<table>
<thead>
<tr>
<th>exid</th>
<th>duration</th>
<th>version</th>
<th>os_cache</th>
<th>scale_factor</th>
<th>out_rows</th>
<th>out_columns</th>
<th>disk_reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2319</td>
<td>17.184</td>
<td>9.4.5</td>
<td>False</td>
<td>8</td>
<td>47989007</td>
<td>8</td>
<td>1416.63157</td>
</tr>
<tr>
<td>2651</td>
<td>2.365</td>
<td>9.4.5</td>
<td>False</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6.500000</td>
</tr>
<tr>
<td>12444</td>
<td>2.457</td>
<td>9.6.0</td>
<td>False</td>
<td>6</td>
<td>36000148</td>
<td>4</td>
<td>0.000000</td>
</tr>
<tr>
<td>5734</td>
<td>9.389</td>
<td>9.4.5</td>
<td>True</td>
<td>4</td>
<td>23996604</td>
<td>10</td>
<td>1272.27272</td>
</tr>
<tr>
<td>4376</td>
<td>0.732</td>
<td>9.4.5</td>
<td>False</td>
<td>1</td>
<td>800665</td>
<td>8</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
Exploring Results

```
from viska import Viska

v = Viska("parameters.yaml")
v.data.head()
```

<table>
<thead>
<tr>
<th>duration</th>
<th>version</th>
<th>os_cache</th>
<th>scale_factor</th>
<th>out_rows</th>
<th>out_columns</th>
<th>disk_reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2319</td>
<td>9.4.5</td>
<td>False</td>
<td>8</td>
<td>47989007</td>
<td>8</td>
<td>1416.63157</td>
</tr>
<tr>
<td>2651</td>
<td>9.4.5</td>
<td>False</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6.500000</td>
</tr>
<tr>
<td>12444</td>
<td>9.6.0</td>
<td>False</td>
<td>6</td>
<td>36000148</td>
<td>4</td>
<td>0.000000</td>
</tr>
<tr>
<td>5734</td>
<td>9.4.5</td>
<td>True</td>
<td>4</td>
<td>23996604</td>
<td>10</td>
<td>1272.27272</td>
</tr>
<tr>
<td>4376</td>
<td>9.4.5</td>
<td>False</td>
<td>1</td>
<td>800665</td>
<td>8</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

This dataset:
~ 15,000 units
~ 300 features
Exploring Results

What is the causal relationship between version and duration?

```
from viska import Viska

v = Viska("parameters.yaml")
v.data.head()
```

<table>
<thead>
<tr>
<th>exid</th>
<th>duration</th>
<th>version</th>
<th>os_cache</th>
<th>scale_factor</th>
<th>out_rows</th>
<th>out_columns</th>
<th>disk_reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2319</td>
<td>17.184</td>
<td>9.4.5</td>
<td>False</td>
<td>8</td>
<td>47989007</td>
<td>8</td>
<td>1416.63157</td>
</tr>
<tr>
<td>2651</td>
<td>2.365</td>
<td>9.4.5</td>
<td>False</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6.500000</td>
</tr>
<tr>
<td>12444</td>
<td>2.457</td>
<td>9.6.0</td>
<td>False</td>
<td>6</td>
<td>36000148</td>
<td>4</td>
<td>0.000000</td>
</tr>
<tr>
<td>5734</td>
<td>9.389</td>
<td>9.4.5</td>
<td>True</td>
<td>4</td>
<td>23996604</td>
<td>10</td>
<td>1272.27272</td>
</tr>
<tr>
<td>4376</td>
<td>0.732</td>
<td>9.4.5</td>
<td>False</td>
<td>1</td>
<td>800665</td>
<td>8</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
Exploring Results

What is the causal relationship between version and duration?

<table>
<thead>
<tr>
<th>duration</th>
<th>version</th>
<th>os_cache</th>
<th>scale_factor</th>
<th>out_rows</th>
<th>out_columns</th>
<th>disk_reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>exid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2319</td>
<td>17.184</td>
<td>9.4.5</td>
<td>False</td>
<td>8</td>
<td>47989007</td>
<td>1416.63157</td>
</tr>
<tr>
<td>2651</td>
<td>2.365</td>
<td>9.4.5</td>
<td>False</td>
<td>2</td>
<td>0</td>
<td>6.500000</td>
</tr>
<tr>
<td>12444</td>
<td>2.457</td>
<td>9.6.0</td>
<td>False</td>
<td>6</td>
<td>36000148</td>
<td>0.000000</td>
</tr>
<tr>
<td>5734</td>
<td>9.389</td>
<td>9.4.5</td>
<td>True</td>
<td>4</td>
<td>23996604</td>
<td>1272.27272</td>
</tr>
<tr>
<td>4376</td>
<td>0.732</td>
<td>9.4.5</td>
<td>False</td>
<td>1</td>
<td>800665</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
Causal Inference

**Goal:** Establish causal relationship between two variables: the **treatment** (cause) and the **outcome** (effect)

**Metric:** *Average Treatment Effect* (ATE):

\[
E(\text{outcome} | \text{treated unit}) - E(\text{outcome} | \text{control unit})
\]

\[
E(\text{duration} | \text{new version}) - E(\text{duration} | \text{old version})
\]
Confounding Effects

Problem: Need to control for confounding effects of other variables

Solution: Consider relevant variables as covariates in causal inference

What is the causal relationship between version and duration?
Confounding Effects

**Problem:** Need to control for confounding effects of other variables

**Solution:** Consider relevant variables as covariates in causal inference

What is the causal relationship between *version* and *duration*?
Results

Average Treatment Effect (ATE):

$$E(\text{outcome} | \text{treated unit, } X=x) - E(\text{outcome} | \text{control unit, } X=x)$$

What is the causal relationship between version and duration?
Results

Average Treatment Effect (ATE):

$$E(\text{outcome} | \text{treated unit, X=x}) - E(\text{outcome} | \text{control unit, X=x})$$

ATE = -4.75 seconds

What is the causal relationship between version and duration?
Covariate Selection

- How do we know which variables to control for?

- clear_os_cache
- in_rows
- out_rows

- version
- duration
Covariate Selection

- How do we know which variables to control for?
Design Decisions as Potential Causes

• Explaining performance differences at level of design decisions

Preliminary results:
Design Decisions as Potential Causes

• Explaining performance differences at level of design decisions

Preliminary results:
• New version of PostgreSQL achieves ~30% higher disk read throughput
• Likely due to newly introduced multi-core support
Viska
(“wisdom” in Icelandic, “to see” in Lojban)

Principled approach to generating, exploring and analyzing performance measurements
Viska
(“wisdom” in Icelandic, “to see” in Lojban)

Principled approach to *generating, exploring* and *analyzing* performance measurements

**Generate**
- Query generator
- Run experiments
- System monitoring
- Data Wrangling

**Explore**

**Analyze**
Viska
(“wisdom” in Icelandic, “to see” in Lojban)

Principled approach to generating, exploring and analyzing performance measurements

**Generate**
- Query generator
- Run experiments
- System monitoring
- Data Wrangling

**Explore**
- Discover trends and correlations
- Python library
- Interactive web app

**Analyze**
Viska
(“wisdom” in Icelandic, “to see” in Lojban)

Principled approach to generating, exploring and analyzing performance measurements

**Generate**
- Query generator
- Run experiments
- System monitoring
- Data Wrangling

**Explore**
- Discover trends and correlations
- Interactive web app
- Python library

**Analyze**
- DAG discovery
- Causal inference
- Hypothesis testing
- Statistical validation
Viska
(“wisdom” in Icelandic, “to see” in Lojban)

Questions & Discussion

**Generate**
- Query generator
- Run experiments
- System monitoring
- Data Wrangling

**Explore**
- Discover trends and correlations
- Interactive web app
- Python library

**Analyze**
- DAG discovery
- Causal inference
- Hypothesis testing
- Statistical validation