UERIOSITY
Toward Curious Data Systems
Abdul Wasay
Data-driven World
Information-driven World
Data Exploration
Modern Data Systems
Modern Data Systems

Data Exploration
How can we improve data exploration?
How can we improve data exploration?

Fast
Interactive
Intuitive
Accurate
How can we improve data exploration?

Queriosity
Vision

Data Canopy
Ongoing
Vision of Queriosity

Can we automate data exploration?
Design Principles

Autonomy

Learning

Usability
Related Work

Improve Performance

SciDB  Fast Aggregates

Redesign Interfaces

dbTouch  Gestural Interfaces

Let the Machine Learn

Information Retrieval  Machine Learning

Data Summarization

Automated Statistician
Autonomy

Data Ranking

Automated IR
Learning

Learn from Data and User

ML to Explore
Usability

Insights at Speed

Interfaces for Big Data
Autonomy

Learning

Usability
How can we improve data exploration?

Queriosity
Vision

Data Canopy
ongoing
Data Canopy

A framework to store and reuse statistics
Statistics and Data Exploration

Maximum Temperature in Boston (F)

Days in March 2016

source: https://www.wundergroud.com
Statistics and Data Exploration

Maximum Temperature in Boston (F)

Days in March 2016

Mean
Mean + std
Mean - std

source: https://www.wunderground.com
Statistics and Data Exploration

Days in March 2016

Cloud Cover

Dew Point Temperature

Temperature

source: https://www.wundergroud.com
Statistics and Data Exploration

Days in March 2016

Cloud Cover

Dew Point Temperature

Temperature

Corr:

source: https://www.wundergroud.com
Statistics and Data Exploration

Statistics:
- Describe variables - Mean, Std
- Summarize major trends - Regression
- Quantify relationships - Corr, Cov
Statistics and Data Exploration

Stat packages - NumPy, SciPy, R
Data systems - MonetDB, Postgres
Hybrids - Psycopg2, MonetDB.R
Statistics and Data Exploration

Stat packages - NumPy, SciPy, R
Data systems - MonetDB, Postgres
Hybrids - Psycopg, MonetDB.R

These approaches are sub-optimal
Incomplete Features

```
<table>
<thead>
<tr>
<th></th>
<th>3.91x10^-3</th>
<th>1.56x10^-2</th>
<th>6.25x10^-2</th>
<th>2.50x10^-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NumPy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postgres</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postgres + NumPy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of rows: 100 Columns
```

```
1x10^2          1x10^3          1x10^4          1x10^5
```

- Calculate in NumPy
- Fetch Data
- MD
- Postgres + NumPy
- NumPy
- Postgres

Graph showing the relationship between total time (in seconds) and the number of rows for different data handling methods with 100 columns.
### Incomplete Features

<table>
<thead>
<tr>
<th>Statistics Support</th>
<th>Data Mgmt.</th>
<th>Performance</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Systems</td>
<td>✓</td>
<td>✓</td>
<td>✕</td>
</tr>
<tr>
<td>Stat Package</td>
<td>✓</td>
<td>✕</td>
<td>✓</td>
</tr>
<tr>
<td>Hybrids</td>
<td>✓</td>
<td>✓</td>
<td>✕</td>
</tr>
</tbody>
</table>
No Reuse

Time (sec)

Postgres
Postgres + NumPy
MonetDB.R
NumPy

Percentage overlap

100 Columns, 10^6 Rows
No Reuse

Correlation

\[ \sum x \] \[ \sum y \] \[ \sum y^2 \] \[ \sum x^2 \] \[ \sum xy \]

Mean

Std

Covariance

Time (sec)

Percentage overlap

Postgres

Postgres + NumPy

MonetDB.R

NumPy

Mean

Std

Covariance
No Reuse

<table>
<thead>
<tr>
<th>Statistics Support</th>
<th>Data Mgmt.</th>
<th>Performance</th>
<th>Flexibility</th>
<th>Store &amp; Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Systems</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Stat Package</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Hybrids</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Data Canopy

Can we reuse computation and data access?
Can we reuse computation and data access?

\[
\frac{1}{n} \sum X
\]

Mean

\[
\left( \frac{1}{n} \sum X^2 \right) - \left( \frac{1}{n} \sum X \right)^2
\]

Variance

\[
\frac{\sum XY - \left( \frac{\sum X}{n} \right) \left( \frac{\sum X}{n} \right)}{\sqrt{\left( \sum X^2 - \left( \frac{\sum X}{n} \right)^2 \right) \left( \sum Y^2 - \left( \frac{\sum Y}{n} \right)^2 \right)}}
\]

Correlation
Can we reuse computation and data access?

\[
\frac{1}{n} \sum X \\
\text{Mean}
\]

\[
\left(\frac{1}{n} \sum X^2\right) - \left(\frac{1}{n} \sum X\right)^2
\]

\[
\text{Variance}
\]

\[
\frac{\sum XY - \left(\frac{\sum X}{n}\right)\left(\frac{\sum X}{n}\right)}{\sqrt{\left(\sum X^2 - \left(\frac{\sum X}{n}\right)^2\right)\left(\sum Y^2 - \left(\frac{\sum Y}{n}\right)^2\right)}}
\]

\[
\text{Correlation}
\]
Can we reuse computation and data access?

**Basic Aggregates**

\[
\frac{1}{n} \sum X \\
\text{Mean}
\]

\[
\left( \frac{1}{n} \sum X^2 \right) - \left( \frac{1}{n} \sum X \right)^2 \\
\text{Variance}
\]

\[
\frac{\sum XY - \left( \frac{\sum X}{n} \right) \left( \frac{\sum X}{n} \right)}{\sqrt{\left( \sum X^2 - \left( \frac{\sum X}{n} \right)^2 \right) \left( \sum Y^2 - \left( \frac{\sum Y}{n} \right)^2 \right)}} \\
\text{Correlation}
\]
Can we reuse computation and data access?

**Overlaps**

\[
\frac{1}{n} \sum X
\]

Mean

\[
\left( \frac{1}{n} \sum X^2 \right) - \left( \frac{1}{n} \sum X \right)^2
\]

Variance

\[
\frac{\sum XY - \frac{(\sum X)(\sum X)}{n}}{\sqrt{\left( \sum X^2 - \frac{(\sum X)^2}{n} \right) \left( \sum Y^2 - \frac{(\sum Y)^2}{n} \right)}}
\]

Correlation

\[
\sum X
\]

\[
\sum X^2
\]

\[
\sum Y
\]

\[
\sum Y^2
\]

\[
\sum XY
\]
Can we reuse computation and data access?

Overlaps

\[
\frac{1}{n} \sum X
\]

\[
\left(\frac{1}{n} \sum X^2\right) - \left(\frac{1}{n} \sum X\right)^2
\]

\[
\frac{\sum XY - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sqrt{\left(\sum X^2 - \left(\frac{\sum X}{n}\right)^2\right)\left(\sum Y^2 - \left(\frac{\sum Y}{n}\right)^2\right)}}
\]
Can we reuse computation and data access?

Composable

\[ \frac{1}{n} \sum X = \frac{1}{n_1} \sum X_1 + \frac{1}{n_2} \sum X_2 + \frac{1}{n_3} \sum X_3 \]
Data Canopy

Can we reuse computation and data access?

Basic Aggregates
Overlaps
Composable
Data Canopy

Store basic aggregates corresponding to a set of statistics and reuse them to compose overlapping statistics.
Data Canopy
Data Canopy
Data Canopy
Data Canopy

\{\text{Mean, Variance}\}

Weekly Mean

\[ t \]
Data Canopy

\{\text{Mean, Variance}\} \quad \text{Weekly Mean}
Data Canopy

\{\text{Mean, Variance}\} \quad \text{Weekly Mean}

\begin{align*}
1 & \quad 2 \\
\hline
\end{align*}
Data Canopy

\{\text{Mean, Variance}\} \quad \text{Weekly Mean}

1 \quad 2 \quad 3

t

\text{Weekly Mean}
Data Canopy

\{\text{Mean, Variance}\}

Weekly Mean

\[ t \]
Data Canopy

\{\text{Mean, Variance}\}

Weekly Mean

\[\sum x, \sum x^2\]

1 2 3 4

\[t\]

\[\sum x, \sum x^2\]

\[\sum x, \sum x^2\]

\[\sum x, \sum x^2\]

\[\sum x, \sum x^2\]

\[\sum x, \sum x^2\]

\[\sum x, \sum x^2\]
Data Canopy

Weekly Mean

\[ t_1 \sum x \sum x^2 \quad t_2 \sum x \sum x^2 \quad t_3 \sum x \sum x^2 \quad t_4 \sum x \sum x^2 \]
Data Canopy

Key Values

<table>
<thead>
<tr>
<th>t2</th>
<th>$\sum x, \sum x^2$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>t1</th>
<th>$\sum x, \sum x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t3</td>
<td>$\sum x, \sum x^2$</td>
</tr>
<tr>
<td>t4</td>
<td>$\sum x, \sum x^2$</td>
</tr>
</tbody>
</table>
Data Canopy

Weekly Mean = \[ \left\{ \frac{1}{7} \text{get(ti)[0]} \right\} \]
Data Canopy

Weekly Mean = \( \frac{1}{7} \sum_{i=0}^{6} \text{get}(t_i)[0] \)

Monthly Mean = \( \frac{1}{31} \sum_{i=0}^{30} \text{get}(t_i)[0] \)

Weekly Variance = \( \frac{1}{7} \sum_{i=0}^{6} [\text{get}(t_i)[1] - \left( \frac{1}{7} \sum_{i=0}^{6} \text{get}(t_i)[0] \right)^2] \)

Monthly std = \( \sqrt{\frac{1}{31} \sum_{i=0}^{30} [\text{get}(t_i)[1] - \left( \frac{1}{31} \sum_{i=0}^{30} \text{get}(t_i)[0] \right)^2]} \)
Data Canopy

Weekly Means = \left\{ \frac{1}{7} \text{get}(ti)[0] \right\}

Monthly Mean = \left\{ \frac{1}{31} \sum \text{get}(ti)[0] \right\}

Weekly Variance = \left\{ \frac{1}{7} \text{get}(ti)[1] - \left( \frac{1}{7} \text{get}(ti)[0] \right)^2 \right\}

Monthly std = \left\{ \sqrt{\frac{1}{31} \sum \text{get}(ti)[1] - \left( \frac{1}{31} \sum \text{get}(ti)[0] \right)^2} \right\}
Data Canopy

\[ \left\{ \frac{1}{31} \sum \text{get}(ti)[0] \right\} \]

Reuse between ranges

\[ \left\{ \frac{1}{7} \text{get}(ti)[1] - \left( \frac{1}{7} \text{get}(ti)[0] \right)^2 \right\} \]

Reuse between statistics

\[ \sqrt{\frac{1}{31} \sum \text{get}(ti)[1] - \left( \frac{1}{31} \sum \text{get}(ti)[0] \right)^2} \]

Both
Data Canopy
Initialization

- Fully Initialized
- Partially Initialized
- Empty – Strictly reuse
Data Structure

A key value data structure

Search by key and key-ranges
Support fast updates

Operations per second

<table>
<thead>
<tr>
<th>Data Structure</th>
<th>Write</th>
<th>Random Reads</th>
<th>Sequential Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorted Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSM tree</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hash table</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SILT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How can we improve data exploration?

Queriosity

Data Canopy

Vision

ongoing
Demo
Demo

In collaboration with Dhruv Gupta
Demo

Data Canopy

In collaboration with Dhruv Gupta
Thank You!