There’s no data like more data

Theo Vassilakis, Founder and CEO
A bit about Theo
Detecting Failures in Utility Pipes
Processing 3-D scans of pipes using SQL pipelines

**Metadata**
- **Key**: Origin_GPS, Direction, Operator
- **Value**: $O_1^A$, $D_1^R$, Bud

**Points**
- **Position**: $P_{1,1}^R$, $P_{1,2}^R$, ..., $P_{1,M}^R$
- **Color**

```sql
CREATE VIEW pipe_lores AS
SELECT
  FLOOR(Position) AS Cell,
  AVG(Position) AS Position,
  AVG(Color) AS Color
FROM pipe
GROUP BY Cell

SELECT detector_UDF(Position, Color)
FROM pipe_lores
```
Multi-Touch Attribution in E-Commerce

- Algorithm speed-up from days to minutes
- Simpler SQL with fewer dependencies
- Fast, flexible deployment in office Windows VM and production Linux

Volume

Velocity

Touchos
Moving Sum of Touches
The Canonical Data Supply Chain
Complex, Slow, Fragmented

- Data pipelines grow organically
- Accidental Complexity → Manual Tasks
- Specialized Tools → Fragmentation of Skills

<table>
<thead>
<tr>
<th>On Prem</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFS/NetApp RDBMS/Oracle</td>
<td>AWS/S3 HDFS/QFS</td>
</tr>
<tr>
<td>Informatica</td>
<td>Hadoop MapReduce Hive</td>
</tr>
<tr>
<td>Teradata Netezza</td>
<td>AWS/Redshift Google/BigQuery</td>
</tr>
<tr>
<td>SAP Hana</td>
<td>AWS/Dynamo</td>
</tr>
<tr>
<td>Excel Tableau Microstrategy</td>
<td>Tableau Online GoodData</td>
</tr>
</tbody>
</table>

- Storage
- Gather ETL
- Analyze Ad-Hoc
- Share In Memory
- Visual Analysis
What Has Changed?

Compute Power
Distributed

90%+ outside of EDW

Networking Speeds
Rapidly Increasing

Terabit speeds a reality

Privacy and Compliance Critical

Inc. in Data Breaches

Big Data
Investments Rising

73%* of organizations invested/plan to invest

* Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years*, Gartner, Inc., Sep 17, 2014

OLAP

From 90/10 to 10/90
How to be ready for anything?

DATA TYPES
- Logs
- Web
- Documents
- Mobile
- Social
- Media

DATA SHAPES
- Flat
- Nested
- Array

DATA STORAGE
- RDBMS
- File System
- No SQL/Hadoop
- OLAP
- SaaS

Speed → Visibility → Generality

Custom → Standard
Imperative → Declarative
One Choice → Hybrid

Visibility

Executives
Business Users
Analysts
Audit/Compliance
Data Scientists

BI Tools
Spreadsheet
SQL Client
Statistical
IDE

Visibility

Statistical
A Simplified Data Supply Chain

On Prem
- NFS/NetApp
- RDBMS/Oracle

Cloud
- AWS/S3
- HDFS/QFS

Data Compute Engine
- Gather
- ETL
- Analyze
- Ad-Hoc
- Share
- In Memory

Integrated, Managed, Scalable

NOT: transactional, OLTP, CEP
Lessons Learnt & Best Practices

SQL illiteracy considered harmful

- SQL Literacy
  - Reads like English
  - Venn diagrams
  - Sufficiently precise to execute
- Compilers and optimizers can really help distribute computation
- Can do JSON in a standard way

```
SELECT customer_name, count(purchase) AS num_purchases
FROM PurchasesTable
WHERE purchase_date >= DATE '2014-01-01'
GROUP BY customer_name
HAVING num_purchases > 3
```

<table>
<thead>
<tr>
<th>customer_name</th>
<th>num_purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edsger W. Dijkstra</td>
<td>4</td>
</tr>
<tr>
<td>Jim Gray</td>
<td>5</td>
</tr>
</tbody>
</table>
### Lessons Learnt & Best Practices

**Binding compute to storage considered harmful**

#### Next generation analytical applications combine data from many sources

<table>
<thead>
<tr>
<th>From inside the organization</th>
<th>From outside the organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(spreadsheets, logs, SaaS applications, etc)</td>
<td>(web, social media, weather, etc)</td>
</tr>
</tbody>
</table>

#### Storage system options are multiplying

- Traditional RDBMS
- NoSQL
- High-throughput file systems (like AWS/S3)
- Hadoop/HDFS

#### Where the data is stored matters less than what you’re doing with it

- Value of the data growing relative to its container
- Value of agility to the organization growing relative to cost of management
Lessons Learnt & Best Practices

Rule-based workflow considered harmful

- Recipes do not age well
- Must reason about dependencies
- Recompute dynamically
- Non-executable workflow glue the worst
- DSL sounds good at first
- Data flow analytics key
Photos meet declarative processing
Photos meet declarative processing
Declarative Media Processing
Generating Mosaics using SQL pipelines

```sql
SELECT
  m.image AS gallery,
  s.image AS source,
  s.x, s.y,
  m.r AS gallery_r,
  m.g AS gallery_g,
  m.b AS gallery_b,
  s.r AS source_r,
  s.g AS source_g,
  s.b AS source_b,
FROM
  GalleryMeans m
JOIN
  SourcePixels s
ON
  s.r=m.r AND
  s.g=m.g AND
  s.b=m.b
```

**SQL Queries**
Video meets declarative processing
Thank You

theov@metanautix.com

Follow

Blog: http://metanautix.com/blog

@Metanautix