Ground: A Data Context Service

Jiaqi Gao
What have we achieved?

- OLTP vs OLAP
- SQL vs noSQL
- Fast scan, fast index
- Compression
- Transactions
- Version
- ...

A DECOUPLED STACK

Big Data
A DECOUPLED STACK

The Good: Agility
The Bad: Dis-integration.
Crisis: Big Metadata

- Lack of standard mechanism to assemble a collective understanding of the origin, scope, usage of the data
- Metadata can be larger than data itself

- Poor productivity
  - Hard to discover whether data exists, how it’s been used previously

- Governance risk
  - Hard to know who access the data, what they do with it, where they put it and so on
Solution: Data Context

- All the information surrounding the use of data
- **ABC of data context**
  - Application Context
  - Behavior Context
  - Change Over Time

- **Ground**: open source data context service
  - Provides central model, API and repository for capturing the data context

- A database assistant
Janet: I bet social media content can predict which customers might cancel their accounts!

Hey Janet! We already paid for a full Gnip feed from Twitter — you can find it here.

By the way: Sue used this following related table and script.
Janet: I bet social media content can predict which customers might cancel their accounts!

Hey Janet! This looks like Twitter JSON. Many people use this script to turn it into a table.

Be careful: When people store outputs from this script, the following fields are often flagged by IT as PII.

BTW, have you tried the sentiment analysis package?
Janet: It looks true! Tweets predict churn!

Sue: share
I wonder if Janet’s sentiment analysis will help with my discount targeting pipeline.

<table>
<thead>
<tr>
<th>TweetId</th>
<th>Text</th>
<th>neg</th>
<th>pos</th>
<th>neut</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>&quot;sad!&quot;</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>&quot;awesome!&quot;</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>57</td>
<td>&quot;go packers!&quot;</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>&quot;fleek!&quot;</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Time passes...

Oh dear. I better call a meeting to introduce better governance on sentiment labeler.

Uh oh, prediction accuracy metrics are down!

FYI: Janet’s wrangling script changed!

<table>
<thead>
<tr>
<th>TweetId</th>
<th>Text</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>&quot;sad!&quot;</td>
<td>sadness</td>
</tr>
<tr>
<td>53</td>
<td>&quot;awesome!&quot;</td>
<td>elation</td>
</tr>
<tr>
<td>57</td>
<td>&quot;go packers!&quot;</td>
<td>sports</td>
</tr>
<tr>
<td>64</td>
<td>&quot;fleek!&quot;</td>
<td>trendy</td>
</tr>
</tbody>
</table>

VERSION HISTORY
12/31/2016 00:00 -800
hash:
6dda491064bcce14f558bf83867b8c247027c423
user: will
WHAT DID CONTEXT ENABLE?

Self-service catalog, wrangling and analytics.

Collective governance of data.

Fueling our model accuracy monitor.

Figuring out which changes introduced the error.

Determining who made the change to help us resolve the issue.

VERSION HISTORY

user: will
Design Requirements

- Model-agnostic
  - Can’t prescribe how metadata is modeled
- Immutable
  - Can’t erase history
- Scalable
  - Log, version history, lineage
- Politically Neutral
  - Interoperate with a wide range of applications
A DAG shows how metadatas are related
RELATIONAL SCHEMA

JSON DOCUMENT
Shows how metadata changes over time
How used the metadata and how
Open Research Questions

Underground

● Workloads
● Common Ground representations
● No-overwrite versioned DB
● Time travel queries: point and trend Graph queries + log analysis
● Consistency

Aboveground

● Content extraction
● Analytic user exhaust
● Socio-technical networks
● Collective governance
● Reproducibility
● Lifecycle of systems that learn
References

- Ground Context: http://www.ground-context.org/