Finding the Needle in an Unstructured Big Data Haystack

Feng Ye
Senior Software Developer
Teradata
Viya Search

- Why search is relevant?
  - Information age with exploding amount of data
    - Business data doubles every 1.2 years per SanDisk
  - Unstructured data more prevalent (Computer World: 80% of all data)
  - Critical step to filter out noise, gain insights and enable analytics

- SAS DNA
  - Every SAS solution will include a search capability

- Highly distributed solution to index and search business data
Leveraging Viya Capabilities

- Cloud-ready, elastic and scalable
- Open analytics coding environment (Java, Python, LUA, SAS)
- Fast, distributed in-memory processing
- Resilient architecture with guaranteed failover
- Out of box in Viya, no need for separate install
Flexible Search Index

- Easy to Populate Records/Documents of Various Fields
  - from database records and observation tables
  - from structured data like HTML, XML - tags suggest fields
  - from unstructured data - free text fills long fields
  - from data mining output - fields take entities of many types

<table>
<thead>
<tr>
<th>title</th>
<th>body_text</th>
<th>date</th>
<th>facets</th>
<th>etc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Top Features

- High performance cloud based indexing/search over big data
- Support for both unstructured (text) and structured data
- Full text searching: tokenized string field search with wildcard and proximity support or exact field match
- Multi-language support (30+)
- Flexible query: nested Boolean logic (AND, OR and NOT) and range filters
- Flexible/high performance aggregations: metric and bucket based aggregations
- Geo support: data tied to location; filtering based on geo shapes
- Incremental indexing: support streaming data input
Key Advantages

- Powerful aggregation: post-search statistics over user-defined record groupings
  - Multiple metrics per aggregation as opposed to one per aggregation
  - Terms aggregation over multiple fields in one shot
  - Sub-second response time of such aggregation for 30 fields over a million documents, on shared cloud
Key Advantages (cont’d)

- Unified search query and filters
- Easy integration with other SAS analytics capabilities
  - Text mining, categorization, sentiment analysis
  - Recommendation on top of search
  - Comprehensive support for 30 languages
- Committed to backward compatibility (SAS standard)
Geo-aware Search

- Data includes geo location information
- Geo filters for various shapes
  - Rectangle
  - Circle
  - Polygon (City, County, State, Country)
- Geo aggregations
  - Geo Distance (circle based)
  - Geo Bounds
Timeline

- Time based aggregation/chronicle trend
- Drill down/zoom in
Facet Search/Aggregation

- Report counts for multiple facets in one shot
- High performance facet report calculation
- Multiple valued fields support
Text Analytics Flow
Future

- A3S integration: REST API
- Dynamic schema support
- Nested document search and aggregation
Demo
Take Aways

- Viya Search offers flexible and powerful search and aggregation capabilities
- It provides geo awareness and enables timeline drawing
- Performance/scalability and fault tolerance