MarkLogic
NoSQL Data Management

MarkLogic's NoSQL Data Management (NDM) Platform

1. **Name:**   Randall Billy

2. **Organization:**   MarkLogic

3. **Contact information, email and phone number:** randall.billy@marklogic.com, 703-638-9041.

4. **Project Description/Narrative:**

Currently, analysts spend the majority of their time simply searching and gathering relevant data, instead of focusing on uncovering new connections or patterns. Responding to NGA's Disparate Data Challenge, MarkLogic presents its NoSQL Data Management (NDM) platform that dramatically expedites analytic discovery while enriching the analytic process. NDM ingested example data and summarily enriched, indexed, associated, and resolved it to show the power of the NoSQL Data Managaement application, including: Data convergence and the ability for foundation, open source and analysis data to reside in the same space. For purposes of the Data Challenge, we provide an example UI to access the power of the NDM platform. However, NDM provides Java and REST based services that multiple applications could call to leverage NGA data assets.   Our demonstration seeks to revolutionize the analyst experience with these objectives:

- Unify enterprise search of highly diverse data environments.
- Enrich and maximize data via value-level metadata.
- Eliminate ontological dependencies and complications.
- Conflate the data pool into a one-object, one-time - object cloud.
- Empower community collaboration and encourage crowdsourcing its communal data.
- Provide flexible and high performance architecture.

The objectives focus on enterprise search capabilities across multiple data sources. For instance, the search term "Tenerife" against MarkLogic's data challenge implementation will return results from GEONAMES, LANSAT, World Factbook (Spain), and PUB. 151 Distances Between Ports sources. The NDM solution then uniquely leverages value-level metadata and 'smart' ontologies to extract the data within data, then rapidly create and relate objects imbued with information pulled from the most up-to-date reporting. This 'one-object,

one-time' capability allows the user to capture patterns and connections that might otherwise remain obscured within the data without performing lengthy and time-consuming analysis.

Background

MarkLogic's NDM solution was implemented as part of NGA's Enterprise Data Layer (EDL) pilot in 2012, proving that all data – regardless of form or format – could be exposed in a single viewer. The underlying principle of NDM is to leave data wherever it resides – but index it all and expose that to consumers. MarkLogic indexes every word or number, word positions, word order, semantic indexing, temporal and spatial indexing, document format, and more.

The EDL Pilot and the follow-on NCL FCMS UDM implementation was based on MarkLogic's Enterprise NoSQL platform, which eliminates time-consuming upfront data modeling while providing the reliability and security enterprise organizations require. All data were tagged with appropriate metadata in EDL and made immediately searchable and discoverable. USSOCOM implemented the same technologies as their SOF Data Layer, maturing NGA's EDL concepts. The latest generation of search and discovery tools built on the MarkLogic organic indexing services are currently being developed for USSOCOM's All-Source Analytic Environment (ASAE). This NDM demonstration highlights the numerous innovations added to MarkLogic since the EDL pilot.

The NoSQL Data Management User Interface

MarkLogic's NDM solution leverages cloud computing and a NoSQL database architecture to implement a highly scalable, powerful, and, dynamic IT environment. This environment provides enterprise search capabilities, unique value/cell level security, flexible ontology modeling schema, and a platform for generating one-object one-time object views.

The UI supports full spectrum conflationary tradecraft employed to support Object-Based Intelligence (OBI) within an intuitive and powerful search and discover capability. The UI also supports crowdsourcing, collaboration, document and object handling, advanced analytics and visualization, and full spectrum production and dissemination.
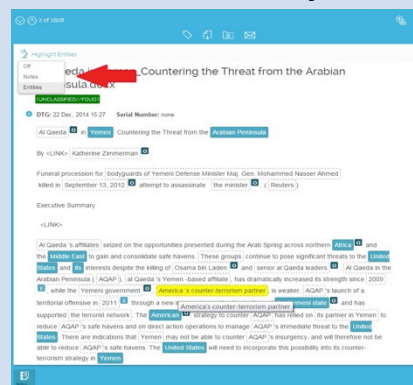
Figure 1: The NoSQL Data Managaement User Interface (NDM UI)
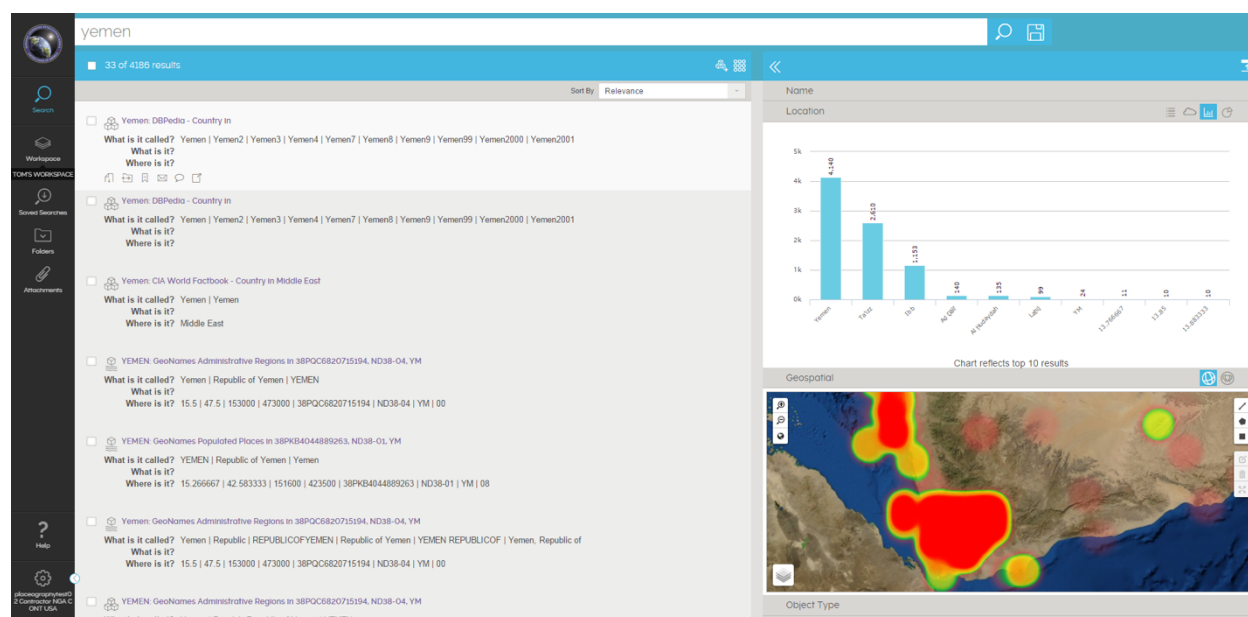
The UI facilitates ingestion, management, and incorporation of multiple disparate data formats to formulate and visualize the one-object one-time instance of that enriched content. The NDM solution's value proposition includes the conjoining and resolving of structured, semi-structured, and unstructured data into single instance objects. The data arises from one or more of the following sources stored natively in MarkLogic:

Table 1: Disparate Data Challenge Ingested Sources in **BOLD**

Unstructured text-based media/Structured Tabular data	• Powered by semantic extraction of unstructured content, the metadata imputed to all content, and attributes of objects or entities found within the underlying repositories. • Flexible data schema and interoperable ontologies allows for interaction between entirely disparate data sets formatted according to divergent classification systems. • **AWS-Phase-1: Word Docs, PPTs, PDFs, XLSs**
Foundation GEOINT	• Foundation GEOINT features become many of the objects in the

	• system. • These data are ingested from a variety of gold-standard content repositories and are systematically dismantled and re-associated with other content to generate, enrich, or conflate other objects. • **Variety of Shapefiles (Hazard and National Map), KML/KMZ, GeoJSON, Airport Ortho Points**
Imagery	• Imagery is appended to our objects as a form of rich media content enrichment. • Imagery establishes baselines and evinces changes to objects or observations of objects over time • **NITF, JPG2000, GeoTiff (LANSAT)**
Structured Tabular data	• Traditional Relational Tables • **GeoDatabase**
Community Collections and Intelligence	• Data can be structured, unstructured, or semi-structured SIGINT, HUMINT, ELINT or other non-GEOINT domain inputs. • **GEONAMES** • **CIA World Fact Book**

Features and associated attribution can be extracted from web services via a library of API's to build new connectors.

Smarter and More Powerful Search

Users can gain new insight into their data using powerful MarkLogic search tools including:

- Faceted search
- Saved search, persistent queries/alerts
- Geospatial Searching
- Temporal Filtering
- Folksonomical Classification
- Relationship and Social Network Filtering
- Saved search, persistent queries and alerts
- Uni- and Multivariate Tag and Word Clouds

MarkLogic's powerful set of search features is additive in nature. Users may begin with a basic keyword search, apply facets for specific types of organizations, temporally bound their search to encompass a period of interest, spatially tie their search to system supplied boundaries such as administrative regions, and then look only for objects or content containing specific relationships of interest.

Figure 2: NDM UI: Search By Bounding Box
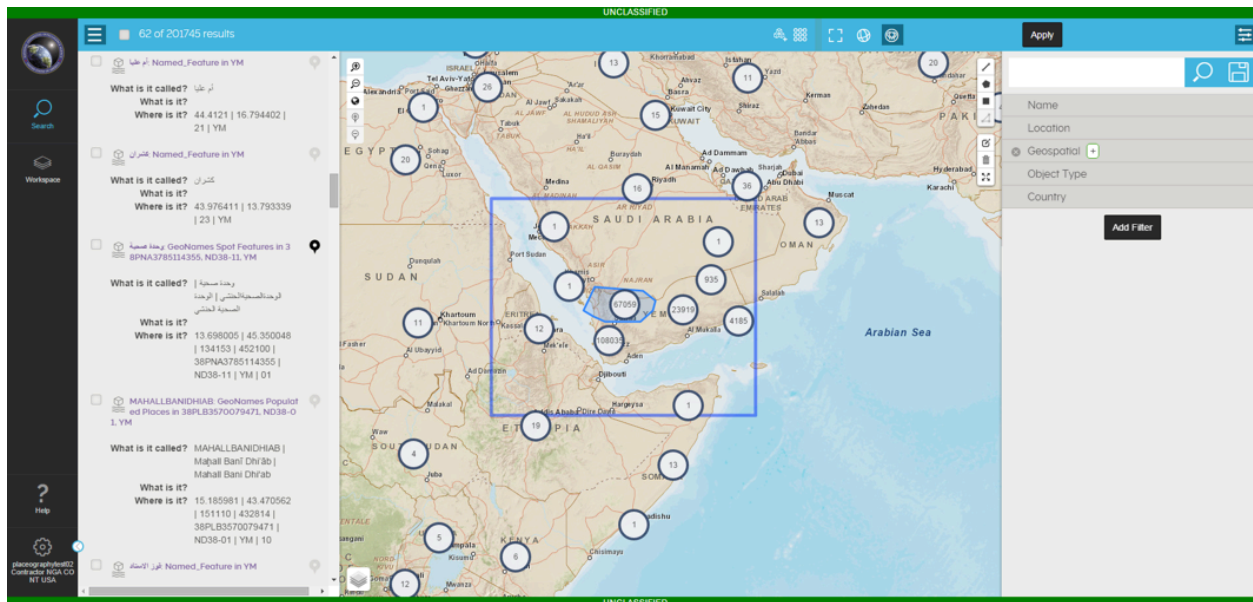
In summary, MarkLogic's NDM solution provides the ideal platform for NGA to store and access disparate data formats.

(946 words)

5. Instructions for accessing our submission:

Link to MarkLogic's demonstration instance at:

https://microdata3.demo.marklogic.com/dataChallenge.

The credentials are data_challenge/password1234!@#$, same as on the bottom of the Data Challenge website used for the S3 bucket.

We did create object views for some of the data that we ingested over three days to provide a look and feel for the inherent capabilities. Here are some examples to explore:

AWS Phase 1 Doc

- Name of Object: 11th Committee Rules.docx
    - Searchable Terms
        - Brian Stack
        - Alexander III of Macedon
        - Alexander the Great

- - - 336 BC
  - Name of Object: EXERCISE, EXERCISE, EXERCISE
    - o Searchable Terms
      - Brian Stack
      - EXERCISE, EXERCISE, EXERCISE
      - Lat: 21.2817
  - Name of Object: Bowditch
    - o Searchable Terms
      - 地理
  - Name of Object: Humanitarian response to the 2004 Indian Ocean earthquake
    - o Searchable Terms
      - £1bn

AWS Phase 1 PDF

  - Name of Object: ADS Chapter 540 USAID Development Experience Information
    - o Searchable Terms
      - Section 540.3.2.10
  - Name of Object: INTERNATIONAL CODE OF SIGNALS
    - o Searchable Terms
      - Radiotelegraphy and radiotelephony
  - Name of Object: PUB. 151 DISTANCES BETWEEN PORTS
    - o Searchable Terms
      - ELEVENTH EDITION 2001

AWS Phase 1 PPT

  - Name of Object: CY 2015 Sample Graphics.pptx
    - o Searchable Terms
      - Tectonic forces or volcanism

This is a list of the specific data available to search for the datasets we've ingested:

LANSAT Data
NTF Data
Geonames
CIA World Factbook (our own contribution, not in the data challenge datasets)
Unstructured data from AWS Phase 1/[doc | xls | ppt | pdf]