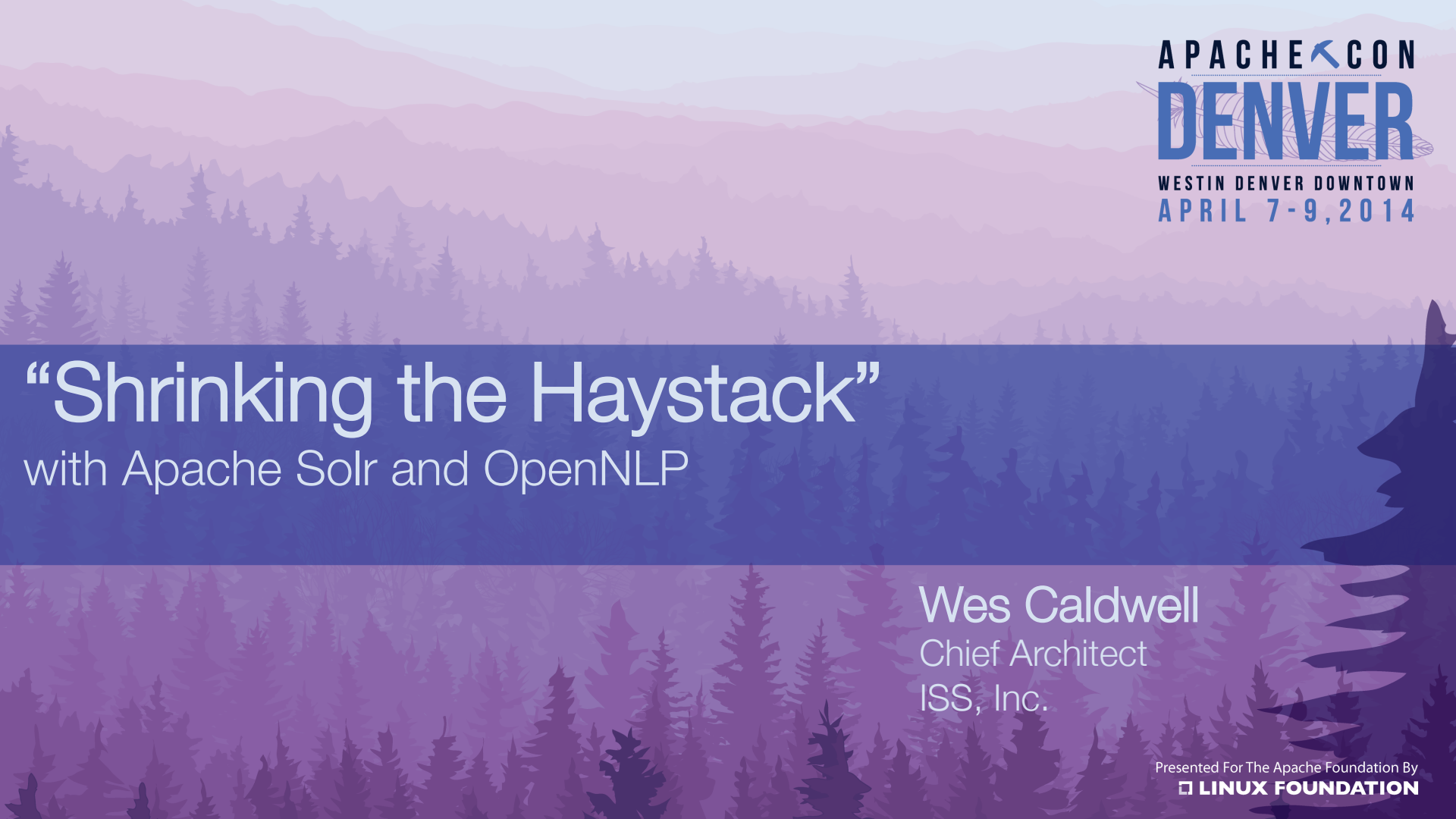



# "Shrinking the Haystack"

## with Apache Solr and OpenNLP

Wes Caldwell
Chief Architect
ISS, Inc.

Presented For The Apache Foundation By
LINUX FOUNDATION

# Topics

- Introduction to ISS, and our customer base
- The data challenges our customers are facing
- Our data processing pipeline (and how Solr and NLP fit in)
- The document processing eco-system
- Additional Solr features that we find useful
- NLP techniques we use
- Why we use multiple NLP techniques and how they complement each other
- Quick demo

# About ISS

- Headquartered in Colorado Springs
  - Other offices located in Washington DC, Hampton VA, Tampa FL, and Rome NY
- Innovative Solutions from “Space to Mud and Everything Between”
  - Sole prime on multiple Air Force Research Labs programs IDIQ
  - Currently Executing More Than 100 Software Development Projects
  - Over 800 employees
  - Strength in Solutions Development and Deployment
- Consistently Recognized as a Leader
  - Recognized as a Deloitte Fast 50 Colorado company and a Deloitte Fast 500 company over eight consecutive years
  - Three-time Inc. Magazine 500 winner
  - 2009 Defense Company of the Year


ISS Intelligent Software Solutions

Numerous awards including Top 100 Contractors by Washington Technology

800+ Employees Headquartered in Colorado Springs

WebTAS®

30,000+ WebTAS users

100+ Programs using WebTAS around the world

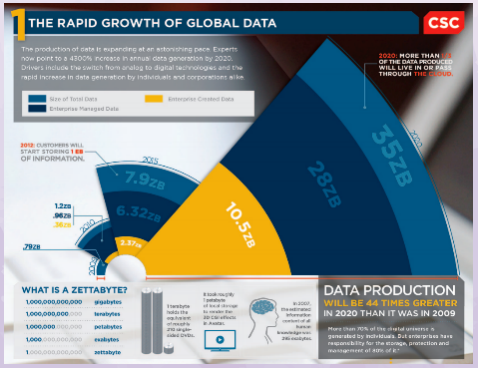
CIDNE

40+ Countries using CIDNE

10,000+ CIDNE queries per day

# The data challenge

- Most electronic information is not relational, but unstructured (textual, binary) or semi-structured (spreadsheet, RSS feed).
  - In 2007, the estimated information content of all human knowledge was 295 exabytes (295 million terabytes)
  - Data production will be 44 times greater in 2020 than in 2009
    - Approximately 35 zetabytes total (35 billion terabytes)
  - A majority of the data produced in the future will be unstructured
- Unstructured data is easily processed by human beings, but is more difficult for machines.
- A tremendous amount of information and knowledge is dormant within unstructured data.



# Our customer's data environment

- Literally thousands of data sources/feeds from a variety of strategic, national, and tactical sources
  - Media (documents, images, etc.)
  - Human Interactions
  - Geospatial
  - Open Source
  - Imagery/Video
  - Many more…

# How our analysts feel

# The need

- Analysts are looking to extract knowledge from the massive heterogeneous data sets, providing “actionable intelligence”
- Search and NLP techniques are key enablers to allow an analyst to reliably search for the information **they know about**, and to assist them in discovering the information **they don’t know about**
- It is critical (especially in tactical environments) to provide tools to the analyst that allow them to “**shrink the haystack**” to a more digestible size, and seed that information into an analytics pipeline, targeted at a particular problem domain (e.g. C-Terrorism, C-Narcotics, etc.)
    - Time-to-live on the relevance of data collected can be very short
    - Its not about finding the needle in the haystack, its about giving a trained analyst the tools to present the most relevant information in a timely manner, allowing them to make an informed decision

# Where our journey led us

# Our approach

## Content Acquisition

Structured Content


Semi-Structured Content


Un-Structured Content

Content Cache (Haystacks)

**Tenets**
- Connector architecture
- Data normalization
- Data staging
- Data Compartmenting (Multiple Haystacks)

## Search/Discovery

Content Index

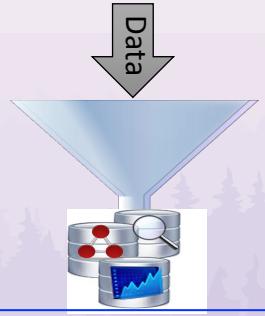
**Tenets**
- Optimized Index of Content for Search and Discovery of Big Data
- Analyst Topics that "Shrink the Haystack"
- Advanced Search Features (Facets, Auto-Complete, Tagging, Comments, etc.)
- Semantic (Synonym) Search based on **pluggable taxonomies**

## Semantic Enrichment


NLP Pipeline
- Categorization
- Named Entity Recognition
- Clustering

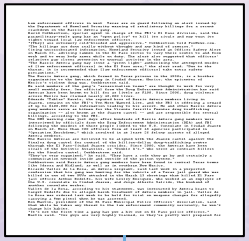
**Tenets**
- **"Domain Spaces"** that support **pluggable** entity recognition and categorization
- Continuous feedback loop that improves the system over time with **analyst input**
- **Lexicon-based** analytics that allows for targeted categorization across corpus of data

## Data Perspectives




**Tenets**
- Data Reduction into focused **"Data Perspectives"**
- Data perspectives stored in **optimized** formats (e.g. Graph, Time Series, Geo, etc.) for the questions being asked
- Leveraging industry-standard parallel processing frameworks for scalable analytics

# Document Processing Eco-System



- Content Management
- Text Extraction
- Named-Entity Recognition
- Geospatial Tagging
- Clustering/ Classification
- Indexing

Apache Jackrabbit

Apache Tika

Unstructured Information Management Architecture
An Apache Project

GeoNames

Apache Solr

hadoop

boilerpipe
Boilerplate Removal and Fulltext Extraction from HTML pages

openNLP

GATE general architecture for text engineering

PostgreSQL

# Additional Solr features we find useful

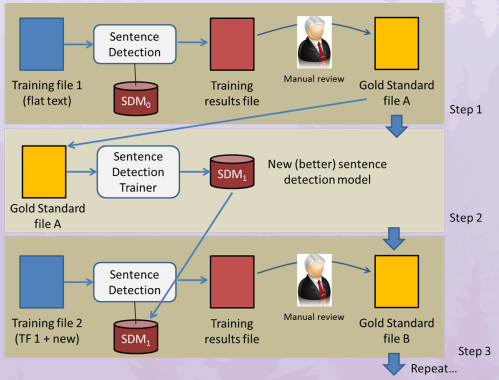
- Synonym (aka “Semantic Search” to us)
  - Allows us to load in pre-defined hierarchical synonym sets(driven by lexicons) to provide search that is tuned for a particular customer domain
    - For example, a search for “weapon” finds various gun types (AK-47, M-16)
  - Currently implemented at index time
  - Simple feature to implement, but has proven very powerful as a “practical analytic”
- Geospatial resolution (used in NLP pipeline)
  - Loaded GeoNames dataset into a separate Solr core
  - Allows for quick lookups in geospatial entity resolution
    - e.g. resolving “Paris” to latitude/longitude based geo-coordinate
  - Can boost based on general rules, or customer-specific ones
    - For example, which “Paris” is it? The one in France or Texas?
      - Population could be the boost parameter that returns Paris, France over Paris, Texas
    - Allows us to easily override for local conditions
      - For example, if a customer wants all geo resolution to be focused in a particular region of the world (i.e. their AOR)

# NLP techniques we use

- Leverage Unstructured Information Management Architecture (UIMA) for NLP pipeline
  - Supports analysis engines for both GATE/Gazetteer and OpenNLP/SML techniques
  - Starting to use UIMA-AS (Asynch) to help in scaling out various pipeline steps
  - Abstracts vendor-specific NLP engine details, hence allowing you to plug in different implementations without much disruption
- GATE/Gazetteer approach
  - Essentially Dictionaries containing key terms used for categorization (facets)
  - Can have n number of “categories” that are generic, as well as customer domain defined
- OpenNLP/Supervised Machine Learning approach
  - “Context aware” models that are trained by data scientists/SMEs
  - Based on probabilistic theory (Maximum Entropy)

# Why use both NLP approaches?

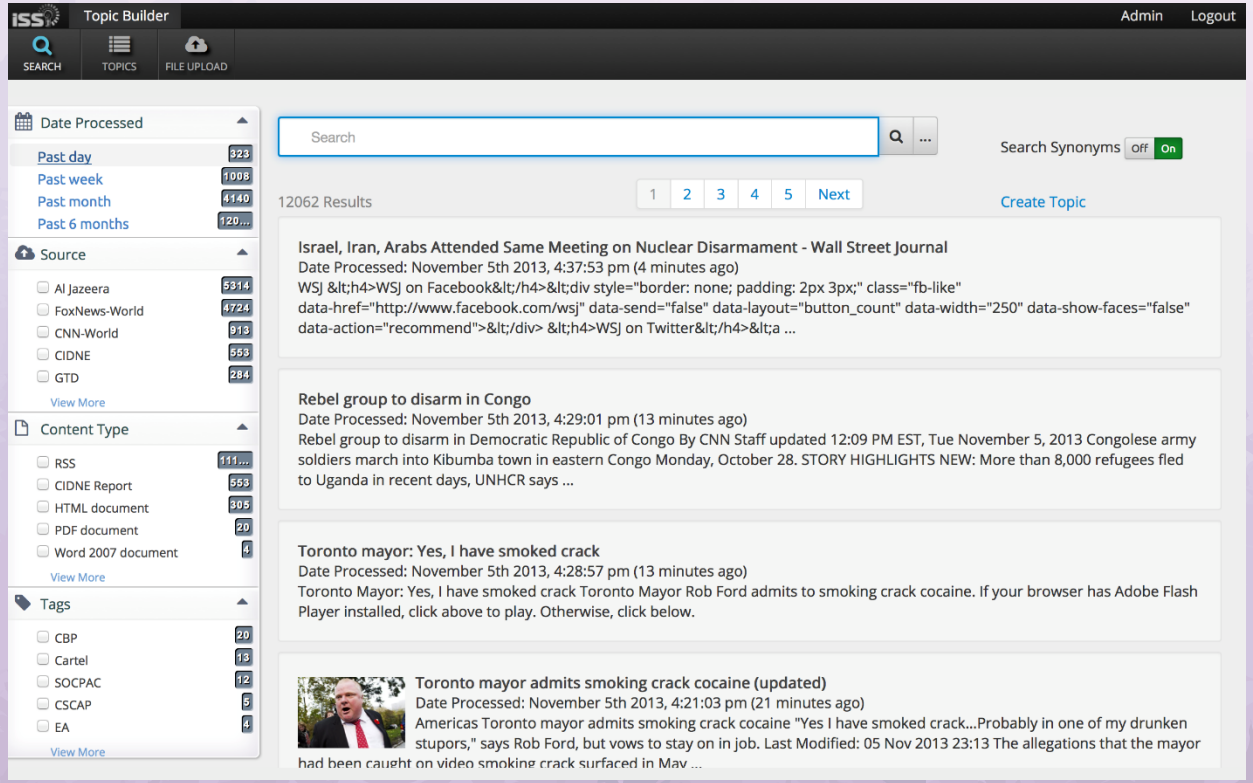
- Both approaches have their pro/cons
- Gazetteer approach
  - Pros
    - Good precision – you are going to find what is important to you
    - Simple for analyst to “tune” - does not require a data scientist
    - Quick and easy to add new categories to a problem domain
  - Cons
    - Only as good as the gazetteer
    - Not context aware
- Supervised Machine Learning approach
  - Pros
    - Once properly trained, good at finding new concepts in context
  - Cons
    - Requires a data scientist/SME to produce quality models
    - Can be tedious to train
- Bottom-line – A combined approach helps you find the things you know are relevant, and also helps you find things that are relevant that you may not know about



# Additional information

- Apache Jackrabbit - http://jackrabbit.apache.org/
- UIMA - http://uima.apache.org/
- GATE - http://gate.ac.uk/
- OpenNLP - http://opennlp.apache.org/
- Boilerpipe - https://code.google.com/p/boilerpipe/
- Apache Tika - http://tika.apache.org/
- Geonames - http://www.geonames.org/

# Demo



# Questions?