# **Developing a Smarter Way to Search – Parsing the online "forest" to find data for your research needs via EDX**

Baker, V.<sup>1,</sup> Bauer, J.<sup>2,</sup> and Rose, K.<sup>2</sup> <sup>1</sup>MATRIC; <sup>2</sup>US Department of Energy National Energy Technology Laboratory, Albany, OR

9000

# Abstract

Scientists conducting data-driven research still spend nearly 80% of their time acquiring, cleaning, and organizing data (CrowdFlower, 2016). Finding relevant, authoritative, and appropriate data remains one of the key obstalces to energy R&D applications. While there is a proliferation of data resources available online, parsing the digital forest to find data "trees" relevant to each user's needs remains a daunting challenge. Although components of new capabilities from big data computing, machine learning, and AI can help address some of these challenges and streamline data management, scientists, and industry still need computing science engineers and domain experts to develop specialized, advanced data computing solutions to harness these capabilities to meet their needs. Addressing this data discovery and acquisition need is crucial to data driven analytics for FE R&D.

Research, scientific, and engineering data resources, including subsurface characterization, modeling, and analytical datasets, are increasingly available through online portals, warehouses, and systems. Developing advanced custom data computing tools to parse data systems (online and network), will improve access and knowledge of data pertinent to FE R&D.

NETL researchers have been developing a novel tool, SmartSearch version 1, to address this need. SmartSearch is a machine learning, online, search tool designed to parse worldwide web for rapid, online, .Zip, and FTP spatial and nonspatial data mining. SmartSearch was recently employed and tested by NETL researchers in the rapid development, 4 months and >2 million resources identified, of a global oil and gas infrastructure, open-source database. SmartSearch v1 is now being incorporated into EDX to support DOE FE user's needs to parse the worldwide web "forest" to rapidly find open source data for a range of end user needs.

Volume, variety, and velocity of data online is growing... exponentially

90%

OF THE NORLD'S DATA

VAS CREATED IN

THE LAST

YEARS

How will you parse your data "trees"

As access to open, authoritative data increases science driven analyses face challenges to efficiently find, integrate and use these resources







**Research & Innovation Center** 



Data is often unstructured, mixed:

- Spatial, contextual
- FTP, WWW, local filesystems, storage area networks, etc.

#### **Convoluted ways to search** for and identify data:

• Hard to identify all the data, i.e., see the whole "Elephant", without falling down the "rabbit hole"



It's a

Fan!

# **NETL's Big Data Discovery Ecosystem (to date)**



from this "forest?"

### Data Discovery Needs

#### Need tools to assist with / automate aspects of data discovery

FROM

TO 2020

We are here

- Parse data silos
- Improve how we use search engines
- Utilize machine learning to correlate relevant information
- Search for data in new ways (<u>e.g., html source</u>)

#### **Need infrastructure capable of processing** billions+ of assets to:

- Extract valuable information
- Understand complex data relationships on a scale previously not possible
- Perform more robust spatio-temporal analyses

# Performs Crawl (Web and/or FTP) of queue and store in Hadoop database

- Aggregates unique http, relative, and ftp links from crawler queue
- Selects, crawls, parses links not previously crawled
- Optionally restrict crawler to specific domains
- Repeat process until threshold (# rows, auqua amptu ata l



Crawler

Table

Select unique links from aggregation of http, relative, ftp links

(Optionally) Filter to restrict to specific domains

Terms	quenes	'search' links		queue empty, etc.)		Apache Tika
	P Spache Spack	Crawler Table	Constant Consta	brocessing / Data Mining: Solr Metastore Manager Workflows ~ Search ~ html:map + Q	Search and/or Contextua	al Cataloging
<u>Ultimately, launch in EDX as a</u> <u>Deep Analysis Recommendation Engine</u>			Grid Results  All (14) / Current (0)  Field Name  version_  content  ftp_links	Image: solution in the solution in		
Related Resources			<ul> <li>hp_inks</li> <li>header</li> <li>html</li> <li>http_links</li> <li>manu_exact</li> <li>payloads</li> <li>redirectedurl</li> <li>rel_links</li> <li>text</li> </ul>	rel_links       http://nipigas.ru/bitrix/components/bitrix/map.yandex.system         header       {null=[HTTP/1.1 200 OK], Expires=[Thu, 19 Nov 1981 08:52:0         html <html xmlns="http://www.w3.org/1999/xhtml"> <head> <link< td="">         html       <html xmlns="http://www.w3.org/1999/xhtml"> <head> <link< td="">         http_links       http://www.individ.ru        version_       1567014083367010304         textcontents       Единое окно обратной связи Раскрытие информации Тенде</link<></head></html></link<></head></html>	ет/ten нефтепереработка Наливные эстакады 00 GMT k rel="st церы Газо	монголия монголи монголи монголи монголи монголи монголи монголи мо
atlasII.pdf Has a 63% match	Atlas-IV-2012.pdf Has a 55% match	Ellergy Data exchange	<ul> <li>text_rev</li> <li>textcontents</li> <li>sourceurl</li> </ul>	<ul> <li>{"sourceurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://www.intergas.kz/ru","redirectedurl":"http://hirbodan.ir/projects/42-</li> </ul>	edirectedur P-mw-cogen	
ATLAS-V-2015.pdf Has a 41% match	Environmental benefit Has a 23% match	<ul> <li>Perform deep contextual analysis</li> <li>Machine learning, natural language</li> </ul>	processing	Building A Big Data Ec FE data driven research re	cosystem for FE R&D equires:	Data Discovery!
File Name       Date         2010atlasIII.pdf       07-01-2014   08:51 AM Eastern		<ul> <li>Generates correlation matches of contextually similar content</li> <li>Being expanded to include spatial and webcrawl assets</li> <li>Implemented using Spark (Scala)</li> <li>Ideal for cluster – RAM, CPU, and bandwidth intensive</li> </ul>		<ul> <li>Lots of data</li> <li>Incorporating different data types &amp; formats,</li> <li>Integrating data from multiple locations (web, local, databases)</li> <li>Traditional Search methods impede our efforts:         <ul> <li>Search engine limits context to a few terms</li> <li>Labor intensive to conduct data searching</li> <li>Even more difficult to find relevant spatial data</li> </ul> </li> </ul>		
Download Stats for all revisions Download Total: 230 Anticipate releasing SmartSearch v1 in next 6 months via EDX						



Disclaimer: This project was funded by the Department of Energy, National Energy Technology Laboratory, an agency of the United States Government, through a support contract with AECOM. Neither the United States Government nor any agency thereof, nor any of their employees, nor AECOM, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.













