Developing a Smarter Way to Search –

Parsing the online "forest" to find data for your research needs via EDX

Baker, V.[1], Bauer, J.[2], and Rose, K.[2] [1]MATRIC; [2]US Department of Energy National Energy Technology Laboratory, Albany, OR

NATIONAL ENERGY TECHNOLOGY LABORATORY

Research & Innovation Center

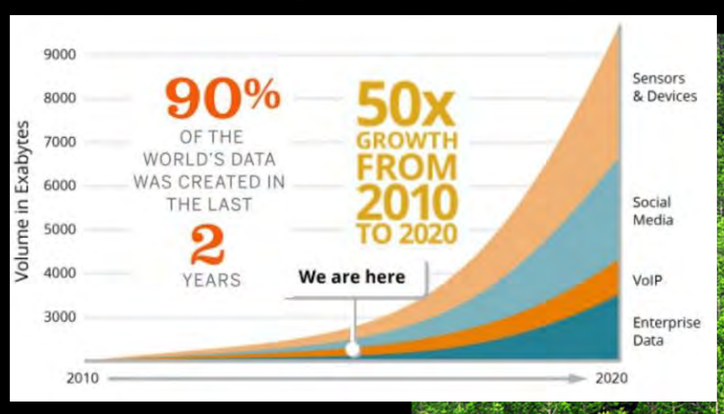
Abstract

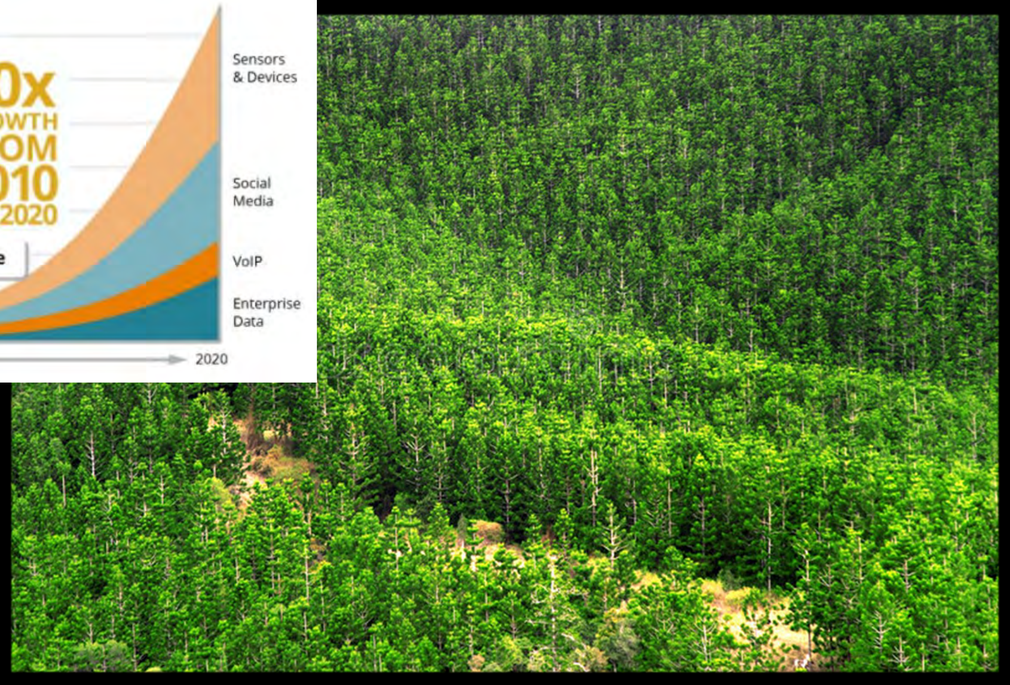
Scientists conducting data-driven research still spend nearly 80% of their time acquiring, cleaning, and organizing data (CrowdFlower, 2016). Finding relevant, authoritative, and appropriate data remains one of the key obstalces to energy R&D applications. While there is a proliferation of data resources 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.

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





- Volume, variety, and velocity of data online is growing… exponentially
- **How will you parse your data "trees" from this "forest?"**

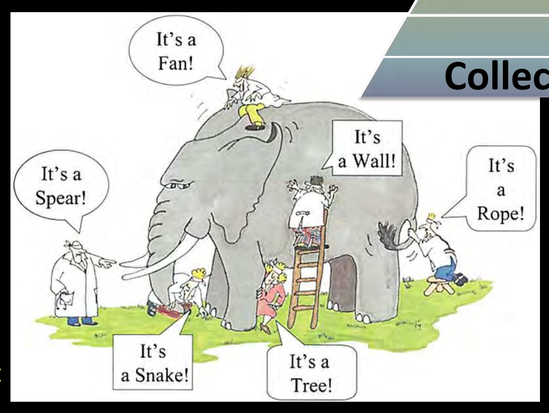
Data Discovery Challenges

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"



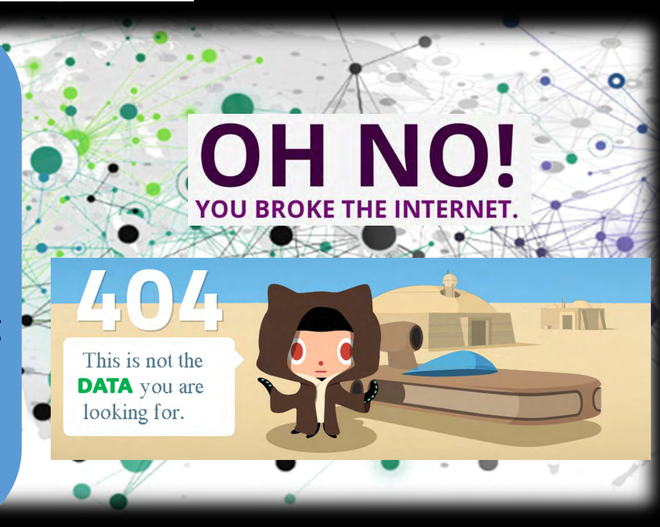
Data Discovery Needs

Need tools to assist with / automate aspects of data discovery

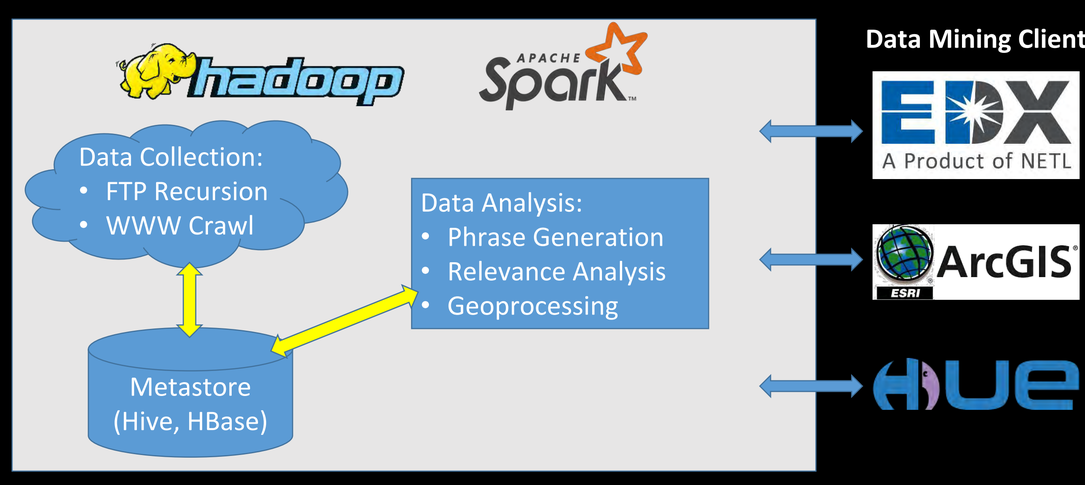
- Parse data silos
- Improve how we use search engines
- Utilize machine learning to correlate relevant information
- Search for data in new ways (e.g., html source)

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



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



Data Mining Clients

EDX A Product of NETL

ArcGIS ESRI

HUE

Query search engine with desired terms to initialize crawler queue

Ingest Categorized Terms → Generate Bing URLs → Execute Bing queries → Parse Result Pages → Crawl Bing relative 'search' links

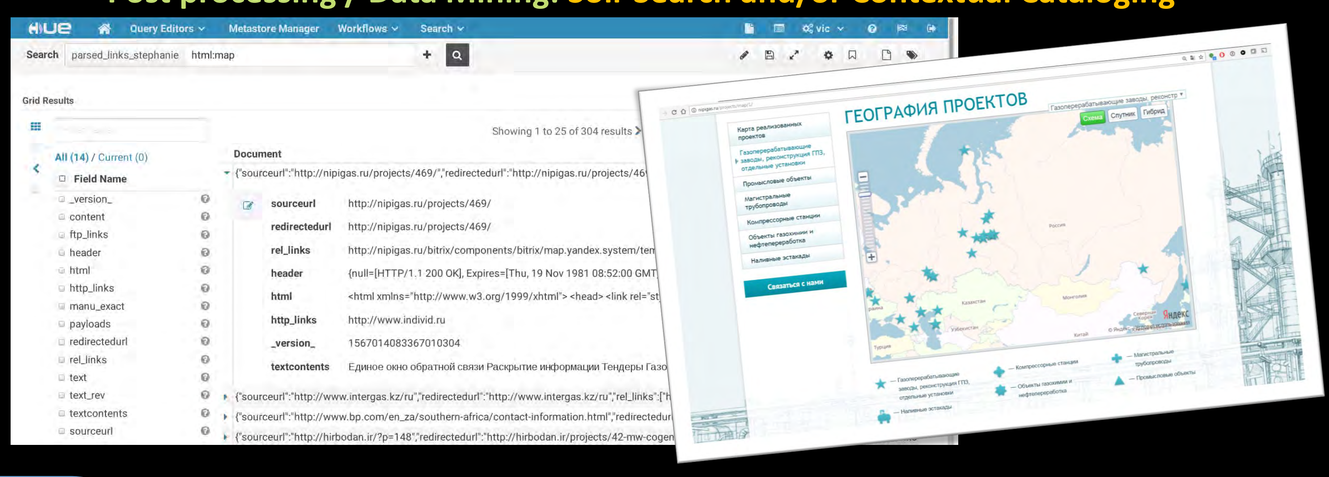
Crawler Table → sourceurl, redirectedurl, header, html, http_links, relative_links, ftp_links, textcontents

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

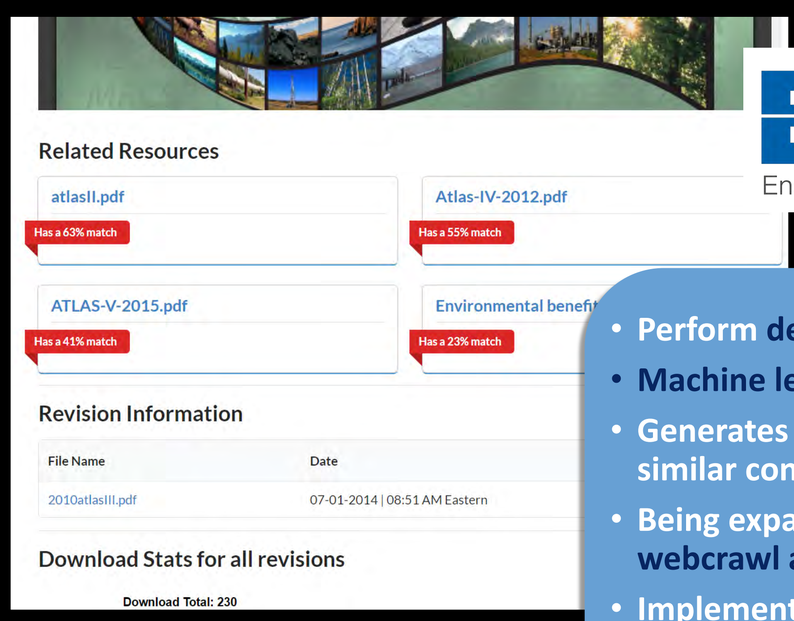
- Aggregates unique http, relative, and ftp links from crawler queue
- Selects, crawls, parses links not previously crawled
- Optionally restrict crawler to specific domains
- Repeat process until threshold (# rows, queue empty, etc.)

Crawler Table → Select unique links from aggregation of http, relative, ftp links → (Optionally) Filter to restrict to specific domains → Open links and parse with Apache Tika

Post processing / Data Mining: Solr Search and/or Contextual Cataloging



Ultimately, launch in EDX as a Deep Analysis Recommendation Engine



EDX™ Energy Data eXchange

- **Perform deep contextual analysis**
- **Machine learning, natural language processing**
- **Generates correlation matches of contextually similar content**
- **Being expanded to include spatial and webcrawl assets**
- **Implemented using Spark (Scala)**
- **Ideal for cluster – RAM, CPU, and bandwidth intensive**

Anticipate releasing SmartSearch v1 in next 6 months via EDX

Building A Big Data Ecosystem for FE R&D Data Discovery!

FE data driven research requires:

- Lots of data
- Incorporating different data types & formats,
- Integrating data from multiple locations (web, local, databases)

Traditional Search methods impede our efforts:

- Search engine limits context to a few terms
- Labor intensive to conduct data searching
- Even more difficult to find relevant spatial data

Acknowledgment: This technical effort was preformed in support of the National Energy Technology Laboratory's ongoing research under the RES contract DE-FE0004000.

U.S. DEPARTMENT OF ENERGY

For more information on NETL's data, and tools visit: https://edx.netl.doe.gov/

POC: Jennifer Bauer
jennifer.bauer@netl.doe.gov

Science & Engineering To Power Our Future