Managing & Analyzing High Volume of Data?

It's Simpler Than You Think!

Suman Biswas – NiyamIT Inc. – IT Solution Architect
Andrew Ditmore – IBM – Program Manager – Risk MAP CDS

ASFPM 2016
The Usual Challenge In Hand

○ Enterprise/organization implementing solutions that collect enormous amount of data

○ Decision makers need to answer key questions in a timely fashion and accurately

○ Analysts are confused:
  ○ Which data set to “use”
  ○ What’s the source of “right” data elements
  ○ How to ensure data is “correct”
  ○ How to put everything together (relationships between data elements) to get to the “answer”
Let’s Adopt A Disciplined Approach

Managing and Analyzing Data – Three Focus Areas

DATA
Layer 1

ANALYTICS
Layer 2

RESULTS
Layer 3

BEST PRACTICES

METHODOLOGIES

FRAMEWORK
It Begins With Data

- Data Governance
- Data Management
- Data Quality

**Data**

**Volume**
- e.g. Flood Disaster Study - 1 sq. mile area – could generate ~8 TB of data

**Velocity**
- Data production will be 44 times greater in 2020 than it was in 2009

**Variety**
- Structured/Unstructured data – weather sensors, social media feed etc.

**Veracity**
- Same data in multiple places – data accuracy?
Data : Key Aspects To Consider

- Master Data Management
  - Identify critical data elements
  - Have a single point of reference – remove duplicates
  - Standardize data and incorporate rules to eliminate incorrect data from entering the system in order to create an authoritative source of master data
- Sometimes, an organization may need to acquire new data elements to answer questions effectively
- Knowledge of the domain is much more important than technical skills
The Rewarding Power Of Data

Gartner Chart

ANALYTICS

e.g. Analysis indicates impending catastrophic event
Analytics: Key Considerations

- Adopt off-the-shelf software or develop custom queries/products
  - SAP Business Objects, IBM Cognos, etc.
- Build models that predict and optimize business outcomes
  - Effective approach to build a model starts with business needs
  - The process to have an effective model may take a few iterations
  - Decide when to have real-time, near real-time or batch data
- Understand your IT infrastructure - physical locations of your data sources
- Analysts & IT staff – both play a pivotal role here
- Keep Big Data factor in mind
Let’s Take Action

○ Reduce/Mitigate Risk
○ Better Decisions
○ Better Processes
Results: Key Considerations

○ Adopt a off-the-shelf software or develop custom solution/dashboard
  ○ Tableau, Qlick etc.
  ○ GIS-based solution is key within FEMA
○ Keep non-IT managers and decision makers in mind as primary stakeholders
○ Results/reports should be light enough for mobile devices
○ Data security is key here too – not everybody should see all the information
Putting It All Together

Data Visualization – Web, Mobile Devices, GIS Apps
(Reporting / Dashboard / Query)

Data Analytics: ad-hoc / pre-compute / notify

Data Transformation / Enrichment Views
(Tailored views of information can be created or recreated on demand from the data lake)

Data Lake
(Secure, distributed repository for all data)

EarthQke  Hurricane  Fire  Tornado  Storm

Other Agencies/Businesses

Infrastructure

Public/Private Clouds

Flexible Infrastructure (Layer 0)

Data (Layer 1)

Analytics (Layer 2)

Results (Layer 3)
Five Key Trends on Data Management & Analytics

- **Analytics standardization is waning**
  - No longer a single solution – organizations are having multiple solutions in terms of Data Management, Analytics (Business Intelligence)

- **Cloud-based data warehousing is on the rise**
  - Cloud-based data warehousing services show the biggest increase in adoption of any information-management category

- **Real-time technology is seeing real gains**
  - Apache Spark, Apache Storm, Splunk, and other options are bringing real-time analysis to system, service, network and mobile scenarios

- **Hadoop and NoSQL adoptions are growing**
  - Factors that are driving NoSQL: faster, more flexible development than achievable with relational databases

- **Data quality concerns are easing**
  - Data is messy -- always has been, always will be. Big-data platforms like Hadoop and NoSQL databases accommodate messiness
FEMA Risk MAP – A Use Case – Coastal Study Data

• The Federal Emergency Management Agency along with a multitude of partners and communities work in cooperation to collect the coastal data needed to produce accurate flood studies.

• It is important to provide transparency to the community and enable them to evaluate the technical approach and the science behind the study.

• In order to update coastal Flood Insurance Rate Maps (FIRMs), data collection is generally conducted along the entire state or county coastline or across an entire lake basin (i.e., for the Great Lakes).

• In addition to providing more accurate flood risk information, these studies make efficient use of limited tax dollars by generating new data that can be used to support other studies on coastal erosion, land use planning, estuaries/wildlife habitat and other coastal topics.

• Under the Risk MAP Program, additional “non-regulatory products” may also be included as data to help communities and property owners plan for resilience.
Coastal Study Data Management – What’s The Issue?

• Currently, this data is stored in various locations, and in some cases may be difficult to obtain since it may have stayed with the partner contracted to collect the data.

• Another challenge has been that this sort of data is typically very large, and in the past this has been a prohibitive factor to creating a single coastal data storage solution.

• This data must be presented in a way that can be easily understood, searchable and accessible.

• The ease of being able to search, access, and obtain the correct coastal data will save time in these future efforts.

• Apply GIS based analytics to help users discover the data in a more meaningful way.

• Initial Data Volume: 260+ TB.

• Same Issue for LiDAR Data Sets.
Coastal Study Data Management – Big Data Upload/Download

• TCP Based File Upload/Download options have limitations

• Implemented FASP™ based solution (UDP) – Aspera Software

• Comparison between traditional HTTP vs. Aspera solution

<table>
<thead>
<tr>
<th></th>
<th>Traditional HTTP Upload/Download Time (in sec)</th>
<th>Aspera Download/Upload Time (in sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.23 GB Download from VA location – Aspera server in VA</td>
<td>1312</td>
<td>456</td>
</tr>
<tr>
<td>1.23 GB Upload from VA location – Aspera server in VA</td>
<td>922</td>
<td>750</td>
</tr>
<tr>
<td>1.23 GB Upload from VA location – Aspera Server in CA</td>
<td>1590</td>
<td>444</td>
</tr>
<tr>
<td>1.23 GB Upload from VA location – Aspera Server in CA</td>
<td>1263</td>
<td>857</td>
</tr>
</tbody>
</table>

• A meaningful structure for the files to be stored
  • Deep down folder tree structure – could have > 20 sub-folders
  • Millions of files

• Validated metadata accompanies each project data upload
Data Is Available In One Place – Need to Search & Retrieve

Solution provides……

- Single system to store, index, catalog, search, discover, display, acquire GEOINT and Multi-INT
- Simultaneously search across local and multiple repositories
- OpenSource Based Solution - ElasticSearch
- Support for over 200 data formats including, geospatial imagery, and NGA standard products
- Powerful natural language and full text search
- Records management functionality
- User defined automated search notifications (standing orders)
- Discoverable with Z39.50 and CSW/OGC
Risk MAP Search Engine

Flood Risk Study Engineering Library

Keyword(s) Search

Go

Advanced Search

Last Updated: Wednesday, 22-June-2016 11:22 AM ET

DHS.gov | FEMA.gov | Privacy Policy | Terms of Use | Accessibility | Site Help | Site Map | Contact Us

FEMA 500 C Street, SW Washington, D.C. 20472 Phone: (202) 566-1600
Risk MAP Search Engine (contd.)
Search & Retrieve Process – Technical Background

Client uses query function to obtain results and use case ID to access relevant data.

Diagram:
- Document oriented NoSQL Partitioned Meta Data Store
- Model, query View 1
- Model, query View 1
- Model, query View 1
- Query Function
- Client
- Existing Data Store
Search & Retrieve Process – High Level View

- Android + Web Server (REST)
- Apple
- Responsive Web Design
- GIS Enabled Viewer
  (Bing, Google)

Diagram:
- Presentation Layer
- Service Layer
- Integration Layer
- File System
- Hadoop Cluster
- NoSQL Mongo DB
- Search Engine
- DB2/Oracle

Mobile
Data Dissemination - Results

- Tableau software was chosen for prototype although many options exist

- Focused on:
  - Easy to read and understand data
  - Navigable by map/geography
  - Putting various ‘measures’ side-by-side
Data Dissemination – Results (contd.)

• Again, selecting an alternate data set from the Table of Contents changes the data visualized but keeps the geography of interest

• The data source stays the same but the tables and maps are updated to show at the new level

• Much like viewing a Region showed a state breakdown, viewing a State shows a County breakdown
THANK YOU!