Accelerate with IBM Storage Webinars

The Free IBM Storage Technical Webinar Series Continues in 2019...

Washington Systems Center – Storage experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

Located in the Accelerate with IBM Storage Blog: https://www.ibm.com/developerworks/mydeveloperworks/blogs/accelerate/?lang=en

Also, check out the WSC YouTube Channel here: https://www.youtube.com/channel/UCNuks0go01_ZrVVF1jgOD6Q

Washington Systems Center - Storage

IBM Spectrum Discover

Content-based Keyword Search and Tagging

Isom Crawford, WSC
Norman Bogard, WSC

Spectrum Discover Update

- Maintenance release, v2.0.1.2 now available
  - Automatic configuration of Content-Search (Tika and agent!)

- Today's focus: New content-search capabilities of IBM Spectrum Discover including
  - what is it? why do I care?
  - how to use regular expressions, and
  - defining content-search policy/policies for metadata extraction.
  - Demonstration with use cases

Content-based Keyword Search & Tagging

What is ”content-based search?”

FEATURE

Out-of-the-box support for content search enables end users to **easily** set up policies to automatically **identify, classify and categorize data**, which could be leveraged for specific business needs

Support for extracting headers, keywords from **hundreds of different file/mime types**

BENEFITS

For the Data Scientist, CIO and the Data Analyst, the ability to curate, extract and gather data containing specific keywords is critical in large scale analytics involving vast amounts of unstructured data.

For the Data Steward and the CIO the ability to find and organize documents based on content greatly helps with their data administration efforts – for example, identifying data that may be subject to specific governance policies and/or compliance regulations.

Content-based Keyword Search

Enables automatic classification of PII & sensitive data

FEATURE

Identifies key fields such as SSN, phone numbers, account numbers and many others to **identify and tag content that contains PII** & Sensitive Data.

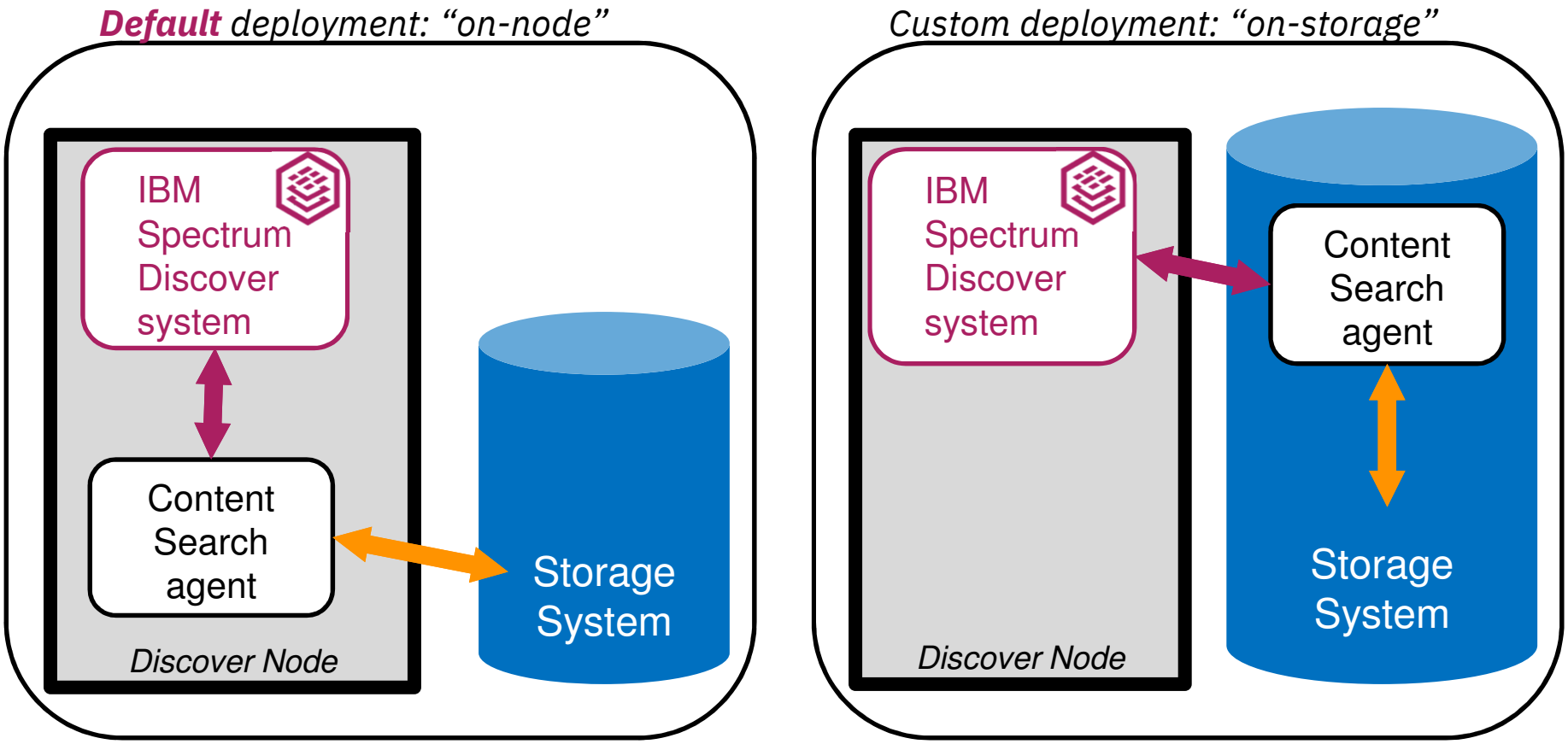
BENEFIT

Automates the **identification and classification** of documents that could potentially contain Personally Identifiable Information (PII) and Sensitive Data.

Out-of-the-box support for content-based data classification enables end users to easily set up policies to automatically identify, classify and categorize data, which could be leveraged for specific business needs

Content Search Functionality in a Nutshell

- Special case of a “Deep Inspection” agent
- Leverages open source content analysis toolkit – Apache™ Tika



The Apache Tika™ toolkit is an ASFv2 licensed open source tool for extracting information from digital documents. Tika allows search engines, content management systems and other applications that work with various kinds of digital documents to detect & extract metadata & content from major file formats.

How does it work??

- Basically the same sequence of steps:
  - ✓scan the data source
  - ✓define tags for the metadata you want
  - ✓define policy to collect the metadata
  - ✓run the policy
  - ✓explore, report, etc.
- One important variation – **regular expression** specification
  - when defining policy, you must select one (or more!) regular expressions
  - what is a “regular expression?” Good question, beyond scope, but here are some good resources:
    - https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference (Good resource from our friends at Microsoft®)
    - https://en.wikipedia.org/wiki/Regular_expression#Syntax (It’s pretty good!)
  - how ***does*** one go about defining a regular expression for their data? Another good question. Header extraction tools frequently exist and one can go directly to Tika too!

Demonstration

Three Use Cases

- Governance / Personally Identifiable Information
  - Searching for credit card numbers, SSN, PINs, etc.
  - Is it where it's supposed to be?
- Genetics
  - Extracting metadata from Variant Call Format (VCF) files
  - Finding format, program information
- Health Care
  - Collecting metadata from Digital Imaging and Communications in Medicine (DICOM) imagery
  - Patient information, image type, etc.

IBM Spectrum Discover Demonstration of Content Search

Enough with the ppt, let’s get to it...

THANK YOU!!

and don't forget:

Free IBM Storage Technical Webinar Series Continues in 2019...

Washington Systems Center – Storage experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

Located in the Accelerate with IBM Storage Blog: https://www.ibm.com/developerworks/mydeveloperworks/blogs/accelerate/?lang=en

Also, check out the WSC YouTube Channel here: https://www.youtube.com/channel/UCNuks0go01_ZrVVF1jgOD6Q

IBM Spectrum Storage