BigDataRevealed Fills the Weaknesses Inherent in Hadoop And makes IoT as easy as 1-2-3 And all with SecureSequester/Encrypt

Hadoop users understand the following barriers must be overcome to have a secure functioning Data Lake.

1. **Hadoop strips Cataloguing/Metadata from files as the file enters the ecosystem.**
   - **Until Cataloguing/Metadata information is rebuilt the Data Lake is of little value.**
2. **Big Data software products lack the sophisticated security mechanisms available with legacy databases.**
   - **As a result, Hadoop Data Lakes are soft targets for intruders to penetrate.**
3. **Locating Personally Identifiable Information and other Sensitive data is difficult and may require many people hours from Data Scientists.**
   - **Therefore, many Data Lakes contain undiscovered Personally Identifiable Information and Sensitive data fields that are vulnerable to attackers.**

Hadoop databases are designed to capture and organize incredible volumes of raw data reaching into the Peta Bytes. A properly built Data Lake can provide a company with a 360% view of it's activities, customers and machinery, but can also supply hackers with the same bounty of information if not properly managed. BigDataRevealed was designed to address each of the above major weakness in Hadoop with limited effort from Data Scientists and Data Management folks.

This is why you want BigDataRevealed on your side to help create a Useful and Protected Data Lake.

- **As Data streams or imports into your Data Lake BigDataRevealed's Intelligent Catalogue will re-create catalogue data and metadata that was stripped away as well as determine the business classification form more precise columnar naming.**

- **Again, as data streams into your Data Lake BDR's Intelligent Catalogue will identify PII and other Sensitive data and Sequester/Encrypt the fields before writing them to HDFS or Hbase. The decryption key is safely stored outside of Hadoop. PII and Sensitive Data are never exposed.**

- **The same processes can be run against 'data at rest' as well as streaming data with little effort.**

- **BDR provides a graphical interface to connect to IoT, and Social Media data feeds directly to your Data Lake, Eliminating the need for Data Scientist to build unique connectors for every data feed you wish to process. Saving many hours of coding and testing while automating the SecureSequester/Encrypt of Personally Identifiable Information.**

BigDataRevealed employs 3 components

BigDataRevealed employs three components to protect streams of data made available by IoT and other devices that write in your data lake

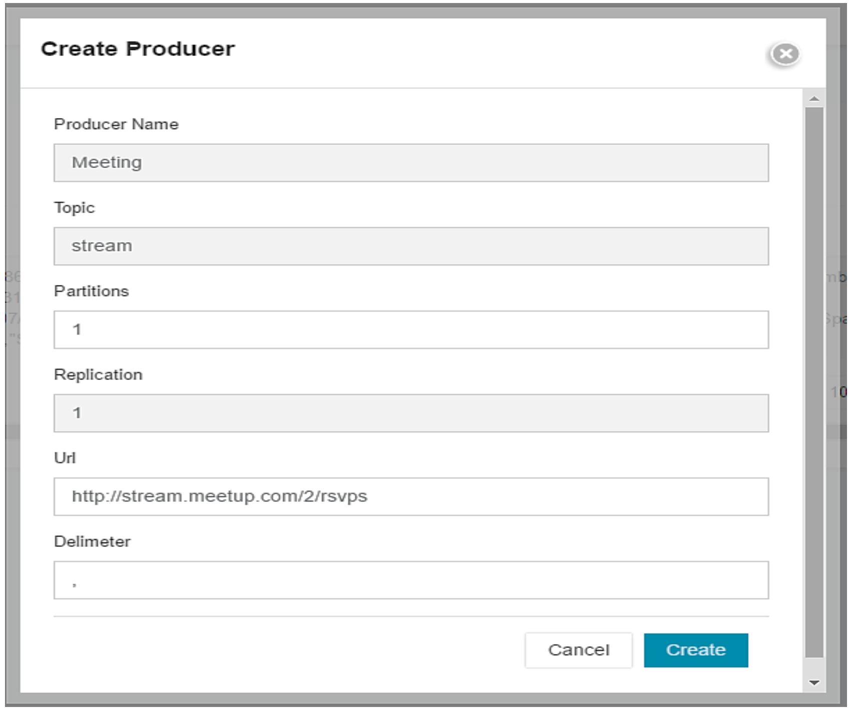
- The producer is the tool that tells BigDataRevealed about the stream of data that you wish to include in the data lake
- The Intelligent Catalogue (SecureSequester Facility) defines what patterns are to be detected as potential Personal Identifiable Information (PII) and which ones have been deemed false positives
- The Intelligent Catalogue (SecureSequester Facility) then takes the configuration information and applies it to streams of information intended for your data lake

I-o-T as easy as 1-2-3

Protection for EU GDPR – US and worldwide Data Protection, Sequester / Encryption

BigDataRevealed employs three components to protect streams of data made available by IOT and other devices that write in your Data Lake

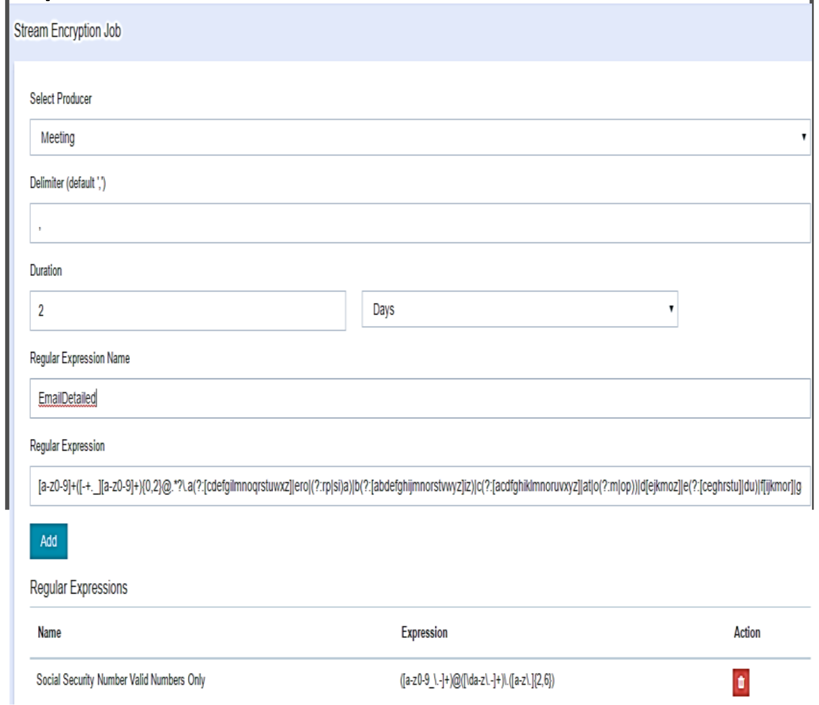
1. Producer is used to register potential streams of information to BigDataRevealed. Connections are automatically generated.



Create Producer

Producer Name: Meeting
Topic: stream
Partitions: 1
Replication: 1
Url: http://stream.meetup.com/2/rsvps
Delimeter: ,

2. The SecureSequester Administrative workbench is used to define PII patterns and known false positives. Set Duration and parameters for the Stream Job.



Stream Encryption Job

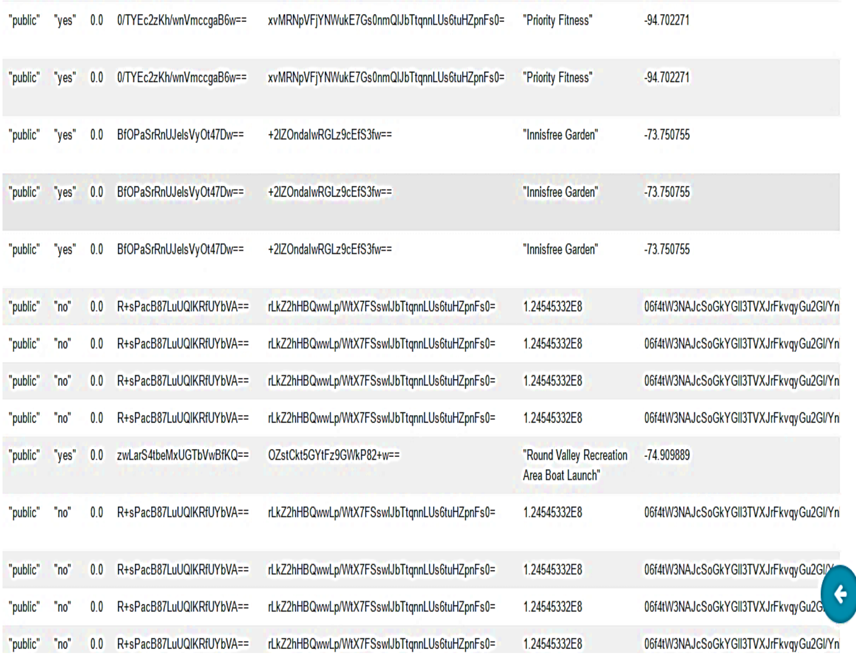
Select Producer: Meeting
Delimiter (default ','): ,
Duration: 2 Days
Regular Expression Name: EmailDetailed
Regular Expression:

```
[a-z0-9]+([-+._][a-z0-9]+){0,2}@.*?\.a(?:[cdefgilmnoqrstuwxz]|ero|(?:rp|si)a)|b(?:[abdefghijmnorstvwyz]iz)|c(?:[acdfghiklmnoruvxyz]|at|o(?:m|op))|d[ejkmoz]|e(?:[ceghrstu]|du)|f[ijkmor]|g
```

Regular Expressions

Name	Expression	Action
Social Security Number Valid Numbers Only	`([a-z0-9_\.-]+)@([\da-z\.-]+)\.([a-z\.]{2,6})`	🗑

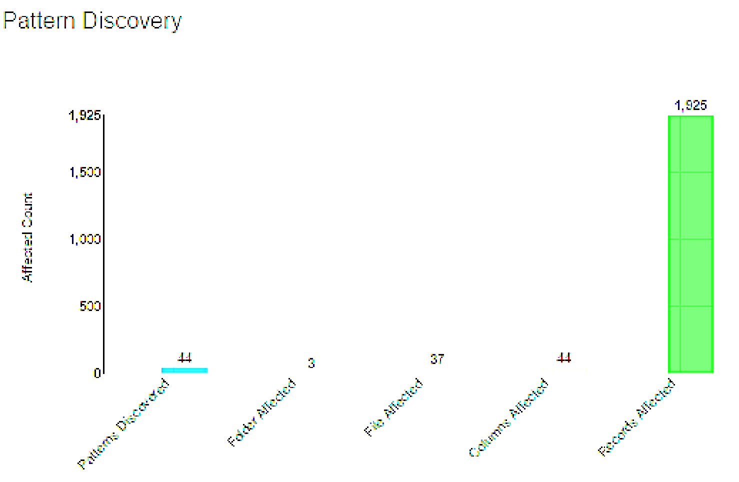
3. SecureSequester will interrogate streams as they are introduced to the data lake, encrypt potential PII, dispatch alerts for review and sequester encrypted source data.

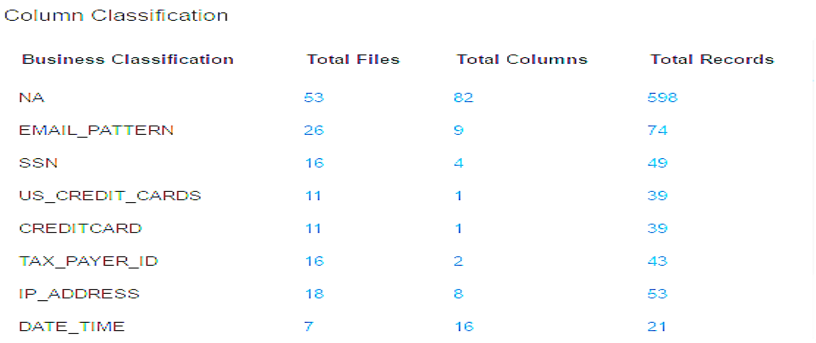


"public"	"yes"	0.0	0/TYEc2zKh/wnVmccgaB6w==	xvMRNpVFjYNWukE7Gs0nmQlJbTtqnnLUs6tuHZpnFs0=	"Priority Fitness"	-94.702271
"public"	"yes"	0.0	0/TYEc2zKh/wnVmccgaB6w==	xvMRNpVFjYNWukE7Gs0nmQlJbTtqnnLUs6tuHZpnFs0=	"Priority Fitness"	-94.702271
"public"	"yes"	0.0	BfOPaSrRnUJelsVyOt47Dw==	+2lZOndalwRGLz9cEfS3fw==	"Innisfree Garden"	-73.750755
"public"	"yes"	0.0	BfOPaSrRnUJelsVyOt47Dw==	+2lZOndalwRGLz9cEfS3fw==	"Innisfree Garden"	-73.750755
"public"	"yes"	0.0	BfOPaSrRnUJelsVyOt47Dw==	+2lZOndalwRGLz9cEfS3fw==	"Innisfree Garden"	-73.750755
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn
"public"	"yes"	0.0	zwLarS4tbeMxUGTbVwBfKQ==	OZstCkt5GYtFz9GWkP82+w==	"Round Valley Recreation Area Boat Launch"	-74.909889
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2G
"public"	"no"	0.0	R+sPacB87LuUQlKRfUYbVA==	rLkZ2hHBQwwLp/WtX7FSswlJbTtqnnLUs6tuHZpnFs0=	1.24545332E8	06f4tW3NAJcSoGkYGlI3TVXJrFkvqyGu2Gl/Yn

Cataloguing / Metadata / Columnar Naming

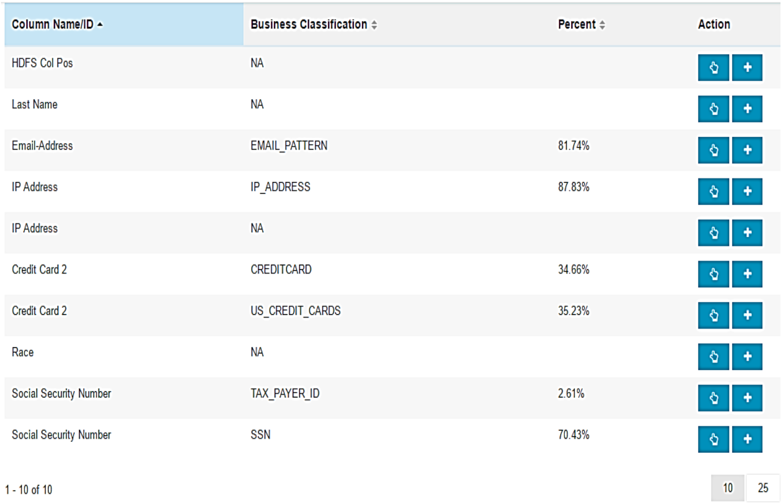
1. Executive Summary of Discovery Process and Patterns Detected.




Column Classification

Business Classification	Total Files	Total Columns	Total Records
NA	53	82	598
EMAIL_PATTERN	26	9	74
SSN	16	4	49
US_CREDIT_CARDS	11	1	39
CREDITCARD	11	1	39
TAX_PAYER_ID	16	2	43
IP_ADDRESS	18	8	53
DATE_TIME	7	16	21

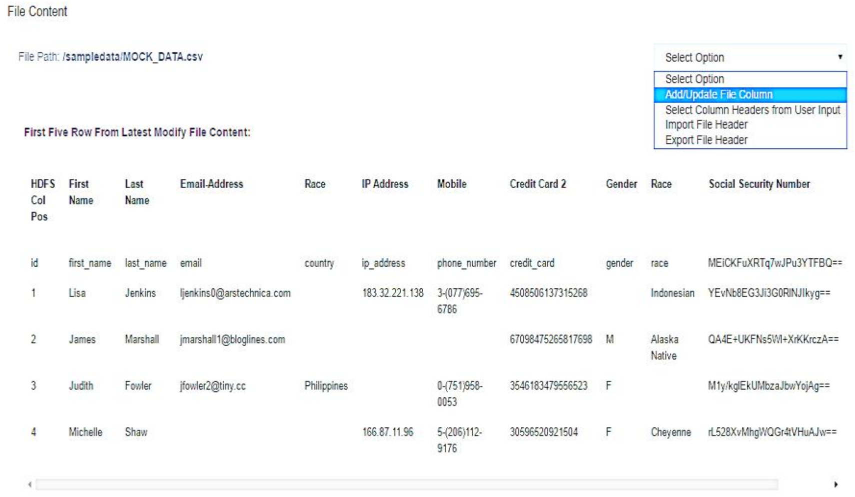
2. BigDataRevealed creating Business Column Classification.



Column Name/ID ▲	Business Classification ⇕	Percent ⇕	Action
HDFS Col Pos	NA		+
Last Name	NA		+
Email-Address	EMAIL_PATTERN	81.74%	+
IP Address	IP_ADDRESS	87.83%	+
IP Address	NA		+
Credit Card 2	CREDITCARD	34.66%	+
Credit Card 2	US_CREDIT_CARDS	35.23%	+
Race	NA		+
Social Security Number	TAX_PAYER_ID	2.61%	+
Social Security Number	SSN	70.43%	+

1 - 10 of 10

3. Creating the File Headers, Catalog Info and Collaborative usage.


File Content

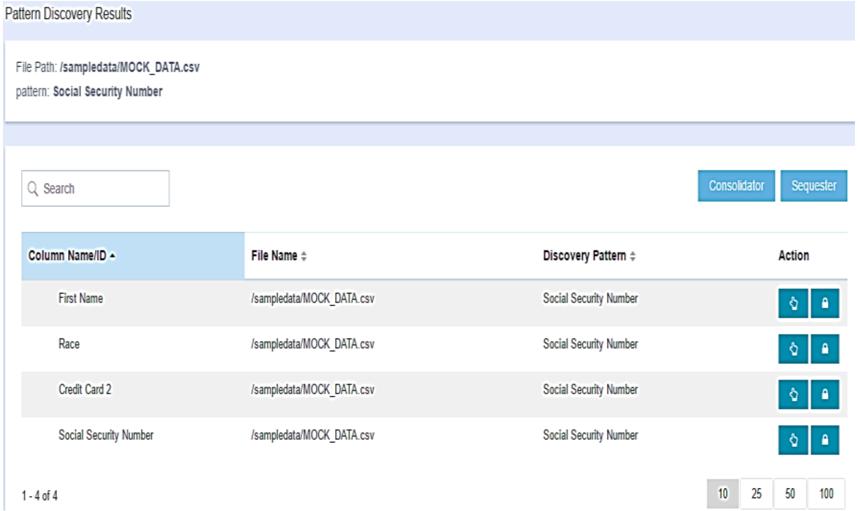
File Path: **/sampledata/MOCK_DATA.csv**

First Five Row From Latest Modify File Content:

HDFS Col Pos	First Name	Last Name	Email-Address	Race	IP Address	Mobile	Credit Card 2	Gender	Race	Social Security Number
id	first_name	last_name	email	country	ip_address	phone_number	credit_card	gender	race	MEiCKFuXRTq7wJPu3YTFBQ==
1	Lisa	Jenkins	ljenkins0@arstechnica.com		183.32.221.138	3-(077)695-6786	4508506137315268		Indonesian	YEvNb8EG3Ji3G0RINJIkyg==
2	James	Marshall	jmarshall1@bloglines.com				67098475265817698	M	Alaska Native	QA4E+UKFNs5WI+XrKKrczA==
3	Judith	Fowler	jfowler2@tiny.cc	Philippines		0-(751)958-0053	3546183479556523	F		M1y/kglEkUMbzaJbwYojAg==
4	Michelle	Shaw			166.87.11.96	5-(206)112-9176	30596520921504	F	Cheyenne	rL528XvMhgWQGr4tVHuAJw==

Pattern Discovery of Private Data then Sequester / Encrypt Protection for EU GDPR

1. View File/Column where Personal Data or Discovered Data was found and select Sequester



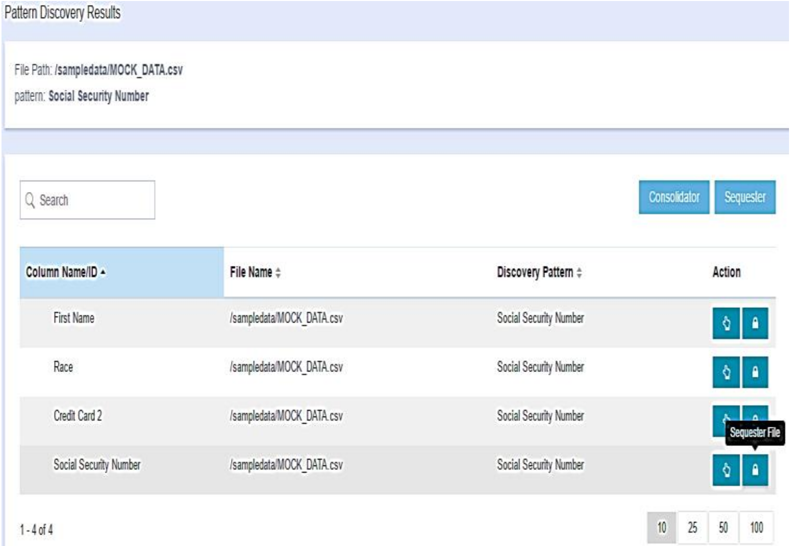
Pattern Discovery Results

File Path: **/sampledata/MOCK_DATA.csv**
pattern: **Social Security Number**

Column Name/ID ▴	File Name ⇕	Discovery Pattern ⇕	Action
First Name	/sampledata/MOCK_DATA.csv	Social Security Number	
Race	/sampledata/MOCK_DATA.csv	Social Security Number	
Credit Card 2	/sampledata/MOCK_DATA.csv	Social Security Number	
Social Security Number	/sampledata/MOCK_DATA.csv	Social Security Number	

1 - 4 of 4

2. Select individual columns by data type to Encrypt, or select entire file to Encrypt. Then Run Process.



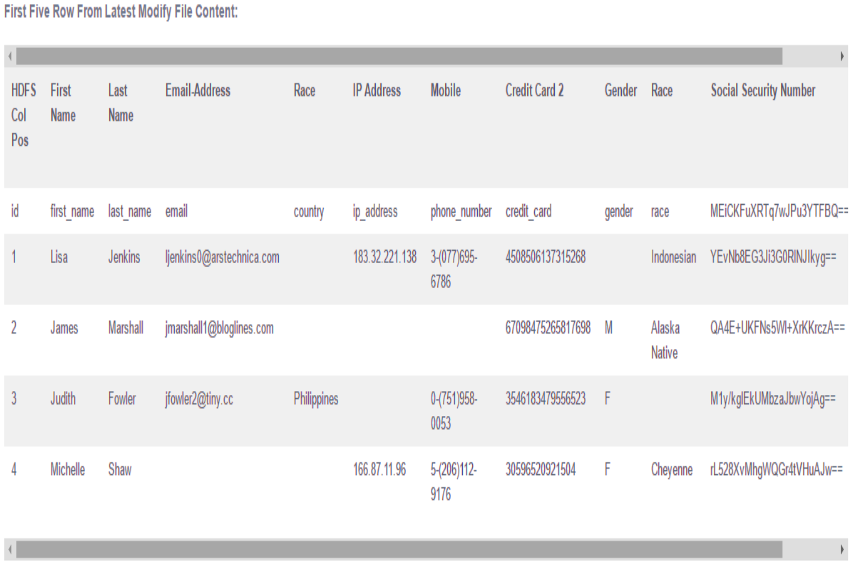
Pattern Discovery Results

File Path: **/sampledata/MOCK_DATA.csv**
pattern: **Social Security Number**

Column Name/ID ▴	File Name ⇕	Discovery Pattern ⇕	Action
First Name	/sampledata/MOCK_DATA.csv	Social Security Number	
Race	/sampledata/MOCK_DATA.csv	Social Security Number	
Credit Card 2	/sampledata/MOCK_DATA.csv	Social Security Number	
Social Security Number	/sampledata/MOCK_DATA.csv	Social Security Number	

1 - 4 of 4

3. View the results of what Columns or if the entire file is encrypted by seeing the actual data.

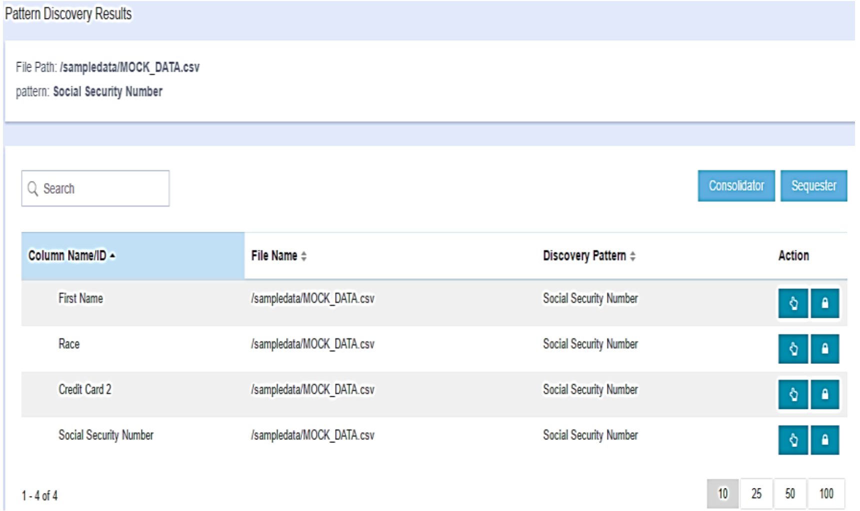


First Five Row From Latest Modify File Content:

HDFS Col Pos	First Name	Last Name	Email-Address	Race	IP Address	Mobile	Credit Card 2	Gender	Race	Social Security Number
id	first_name	last_name	email	country	ip_address	phone_number	credit_card	gender	race	MEiCKFuXRTq7wJPu3YTFBQ==
1	Lisa	Jenkins	ljenkins0@arstechnica.com		183.32.221.138	3-(077)695-6786	4508506137315268		Indonesian	YEvNb8EG3Ji3G0RlNJIkyg==
2	James	Marshall	jmarshall1@bloglines.com				67098475265817698	M	Alaska Native	QA4E+UKFNs5Wl+XrKKrczA==
3	Judith	Fowler	jfowler2@tiny.cc	Philippines		0-(751)958-0053	3546183479556523	F		M1y/kglEkUMbzaJbwYojAg==
4	Michelle	Shaw			166.87.11.96	5-(206)112-9176	30596520921504	F	Cheyenne	rL528XvMhgWQGr4tVHuAJw==

Pattern Discovery of Private Data then Consolidate into one Folder for further Analysis and Remediation

1. View File/Column where Personal Data or Discovered Sensitive Data was found and select Sequester / Encrypt the sensitive data.



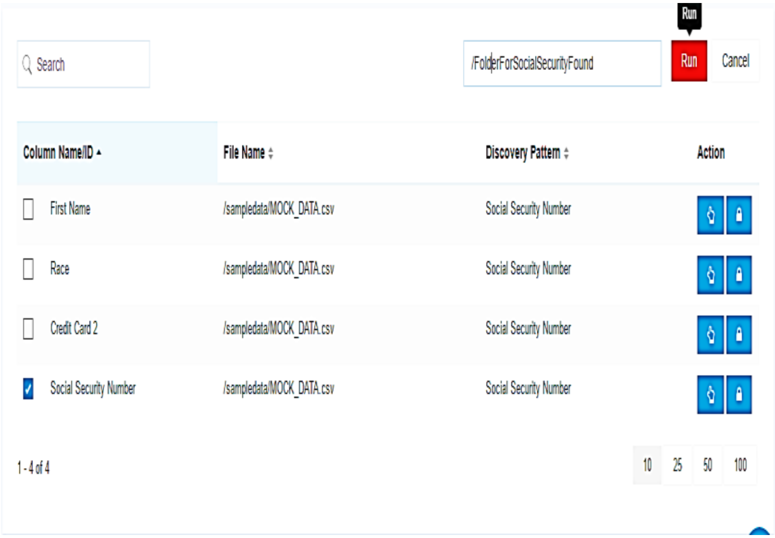
Pattern Discovery Results

File Path: **/sampledata/MOCK_DATA.csv**
pattern: **Social Security Number**

Column Name/ID ▲	File Name ⇕	Discovery Pattern ⇕	Action
First Name	/sampledata/MOCK_DATA.csv	Social Security Number	
Race	/sampledata/MOCK_DATA.csv	Social Security Number	
Credit Card 2	/sampledata/MOCK_DATA.csv	Social Security Number	
Social Security Number	/sampledata/MOCK_DATA.csv	Social Security Number	

1 - 4 of 4

2. Select files to be copied into a new folder containing like data for further analytics, and remediation.

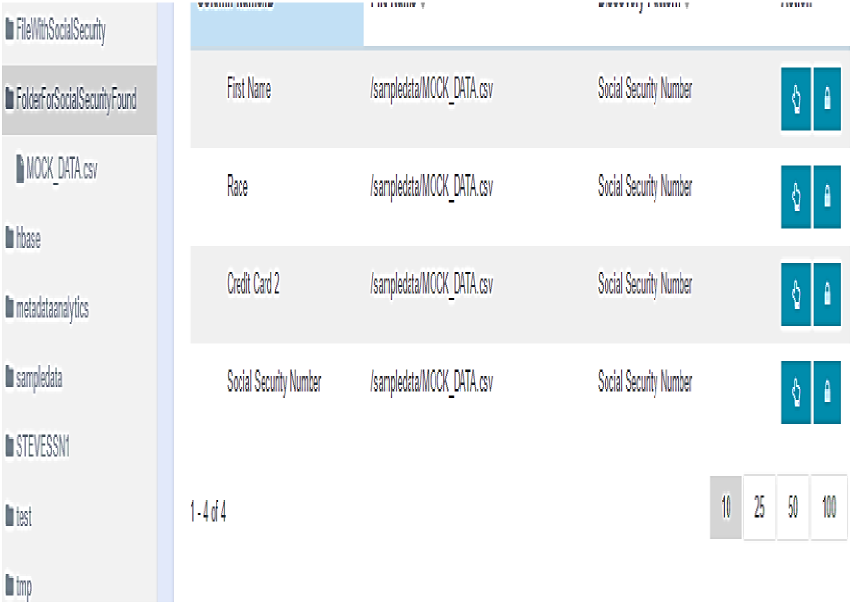


/FolderForSocialSecurityFound

	Column Name/ID ▲	File Name ⇕	Discovery Pattern ⇕	Action
☐	First Name	/sampledata/MOCK_DATA.csv	Social Security Number	
☐	Race	/sampledata/MOCK_DATA.csv	Social Security Number	
☐	Credit Card 2	/sampledata/MOCK_DATA.csv	Social Security Number	
☑	Social Security Number	/sampledata/MOCK_DATA.csv	Social Security Number	

1 - 4 of 4

3. View the results of files that where written into the New Folder as per number 2.



- FileWithSocialSecurity
- FolderForSocialSecurityFound
  - MOCK_DATA.csv
- hbase
- metadataanalytics
- sampledata
- STEVESSN1
- test
- tmp

First Name	/sampledata/MOCK_DATA.csv	Social Security Number	
Race	/sampledata/MOCK_DATA.csv	Social Security Number	
Credit Card 2	/sampledata/MOCK_DATA.csv	Social Security Number	
Social Security Number	/sampledata/MOCK_DATA.csv	Social Security Number	

1 - 4 of 4

info@bigdatarevealed.com

847-791-7838