# Introduction to Big Data

## ECAP456

Dr. Rajni Bhalla

Associate Professor

# Learning Outcomes

After this lecture, you will be able to

- Learn about data ingestion
- Learn about uploading data

# Introduction

Data ingestion in Splunk happens through the Add Data feature which is part of the search and reporting app. After logging in, the Splunk interface home screen shows the Add Data icon

# Introduction

## About uploading data

When you add data to your Splunk deployment, the data is processed and transformed into a series of individual events that you can view, search, and analyze

# Introduction

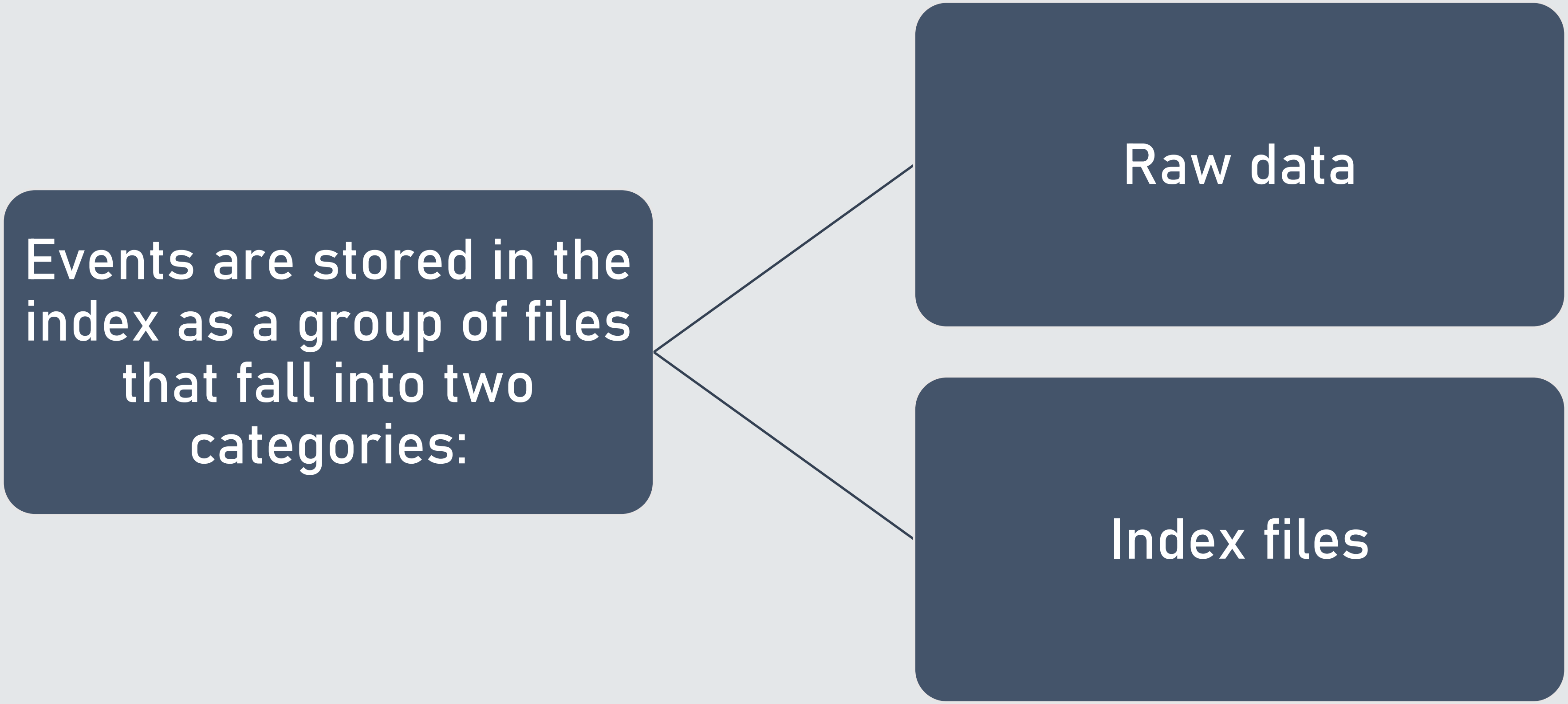
## What kind of data?

| Data source |
|---|
| Files and directories |
| Network events |
| IT Operations |
| Cloud services |
| Database services |
| Security services |
| Virtualization services |
| Application servers |
| Windows sources |
| Other sources |

# Where is the data stored?

- The process of transforming the data is called indexing.
- During indexing, the incoming data is processed to enable fast searching and analysis. The processed results are stored in the index as events.
- The index is a flat file repository for the data. The index resides on the computer where you access your Splunk deployment.

# Where is the data stored?



Events are stored in the index as a group of files that fall into two categories:

Raw data

Index files

# Where is the data stored?

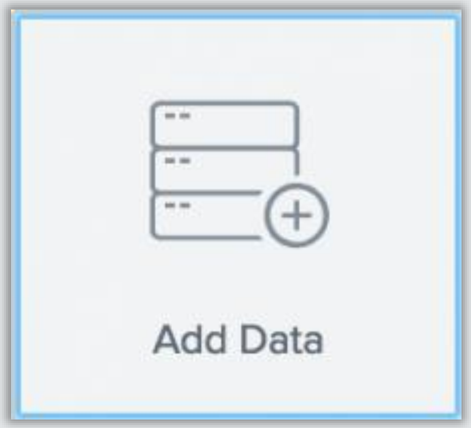
Buckets

Main

Splunk Instance

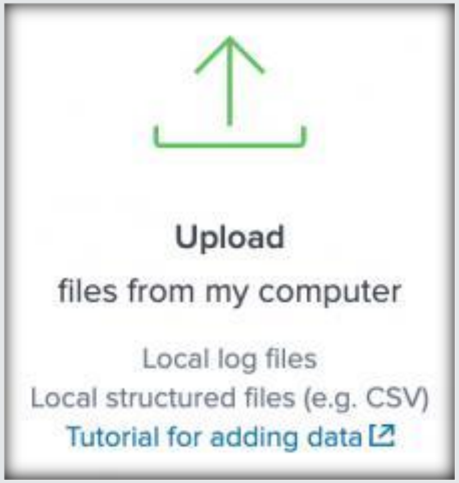
# Use the Add Data wizard

- If there is a Welcome window displayed, close that window.
- Click Settings > Add Data.



# Use the Add Data wizard

At the bottom of the window, click Upload. There are other options for adding data, but for this tutorial you will upload the data files.



**Upload**

files from my computer

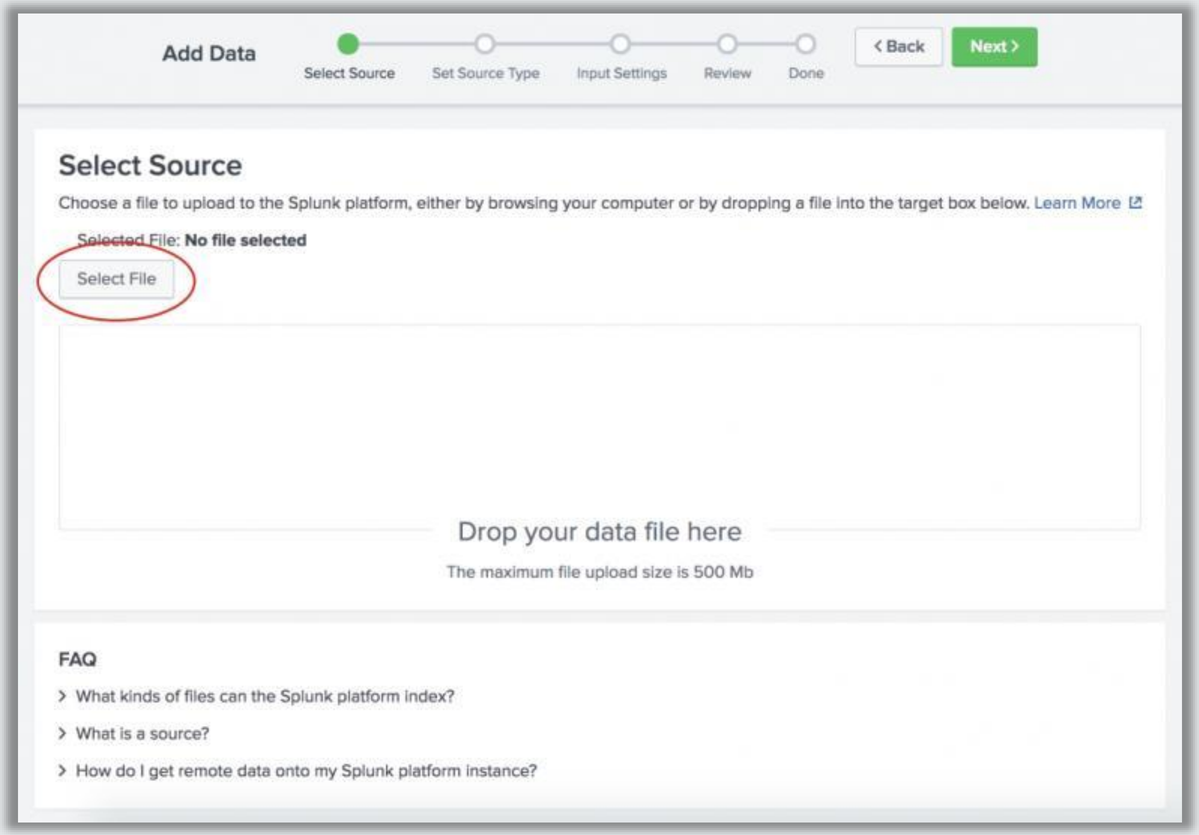
Local log files

Local structured files (e.g. CSV)

Tutorial for adding data

# Use the Add Data wizard

## Under Select Source, click Select File..



Select Source

Choose a file to upload to the Splunk platform, either by browsing your computer or by dropping a file into the target box below. Learn More

Selected File: No file selected

The maximum file upload size is 500 Mb

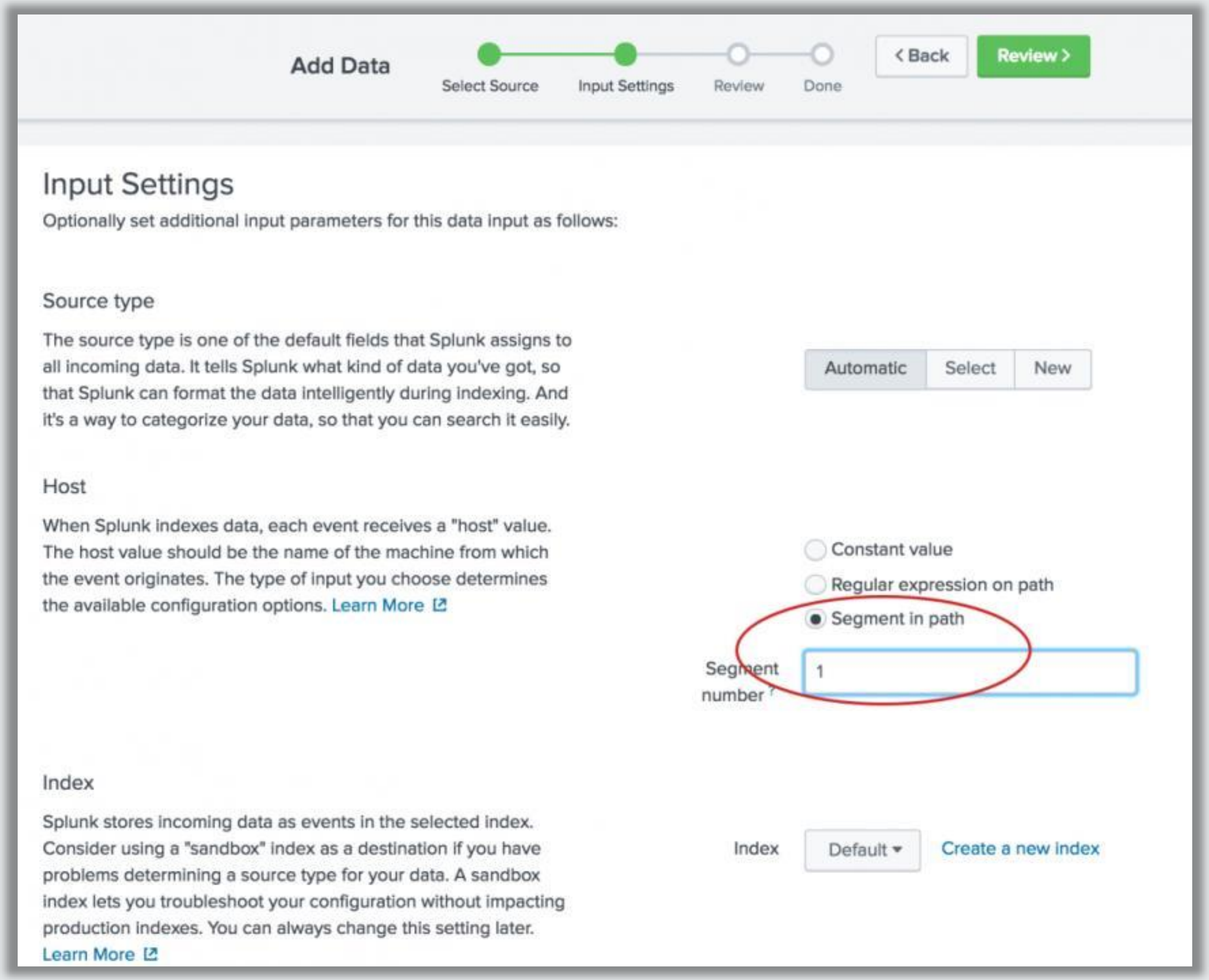
FAQ

- What kinds of files can the Splunk platform index?
- What is a source?
- How do I get remote data onto my Splunk platform instance?

# Use the Add Data wizard

- In your download directory, select zip file and click Open.
- Click Next to continue to Input Settings.
- Under Input Settings, you can override the default settings for Host, Source type, and Index. Because this tutorial uses a ZIP file, you are going to modify the Host setting to assign the host values by using a portion of the path name for the files included in the ZIP file. The setting that you specify depends whether you are using Splunk Cloud or Splunk Enterprise, and on the operating system that you are using.

# Use the Add Data wizard



## Input Settings

Optionally set additional input parameters for this data input as follows:

### Source type

The source type is one of the default fields that Splunk assigns to all incoming data. It tells Splunk what kind of data you've got, so that Splunk can format the data intelligently during indexing. And it's a way to categorize your data, so that you can search it easily.

### Host

When Splunk indexes data, each event receives a "host" value. The host value should be the name of the machine from which the event originates. The type of input you choose determines the available configuration options. Learn More

○ Constant value

○ Regular expression on path

● Segment in path

Segment number: 1

### Index

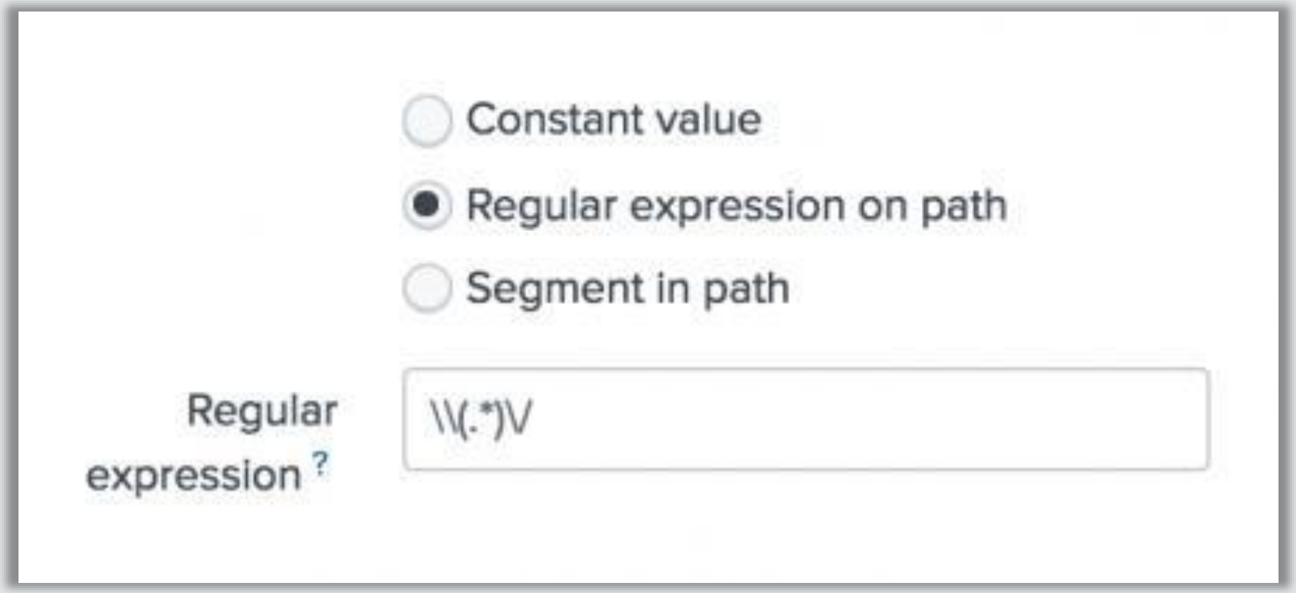
Splunk stores incoming data as events in the selected index. Consider using a "sandbox" index as a destination if you have problems determining a source type for your data. A sandbox index lets you troubleshoot your configuration without impacting production indexes. You can always change this setting later. Learn More

Index: Default

# Use the Add Data wizard

a. Select Regular expression on path.

b. Type \\(.*)\/ for the regex to extract the host values from the path.



○ Constant value

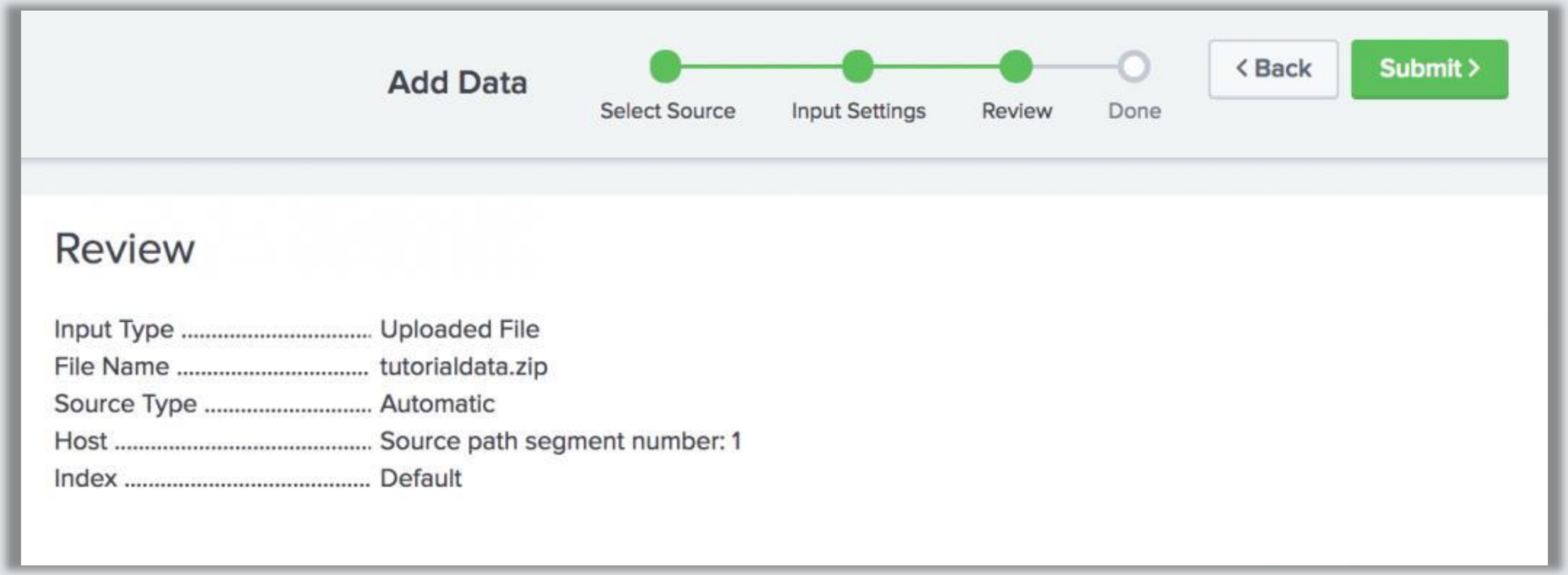
● Regular expression on path

○ Segment in path

Regular expression ? \\(.*)\/

# Use the Add Data wizard

Click Review. The following screen appears where you can review your input settings.

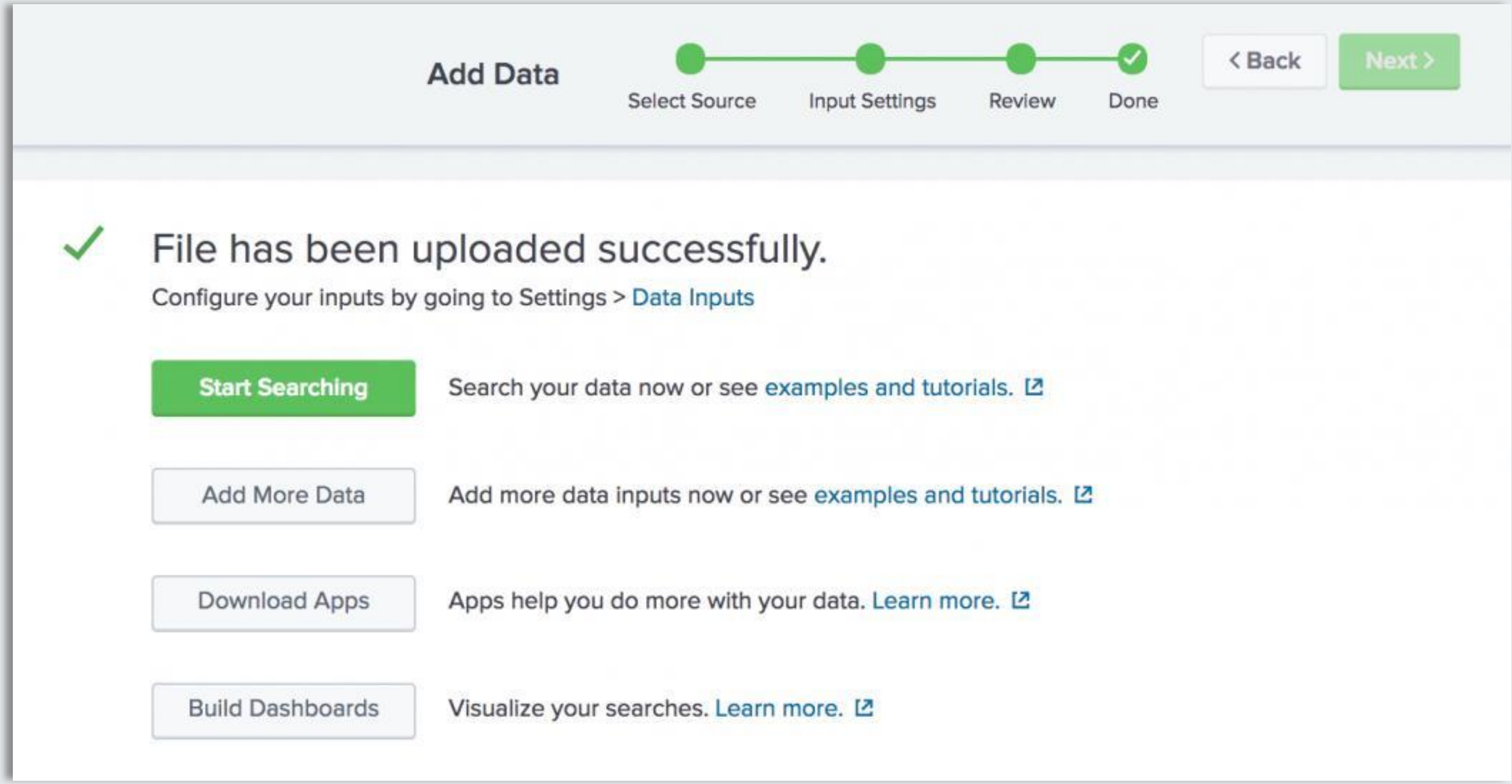


Review

Input Type ................................ Uploaded File
File Name ................................ tutorialdata.zip
Source Type ............................ Automatic
Host ........................................... Source path segment number: 1
Index ......................................... Default

# Use the Add Data wizard

Click Submit to add the data.



✓ File has been uploaded successfully.

Configure your inputs by going to Settings > Data Inputs

Search your data now or see examples and tutorials.

Add more data inputs now or see examples and tutorials.

Apps help you do more with your data. Learn more.

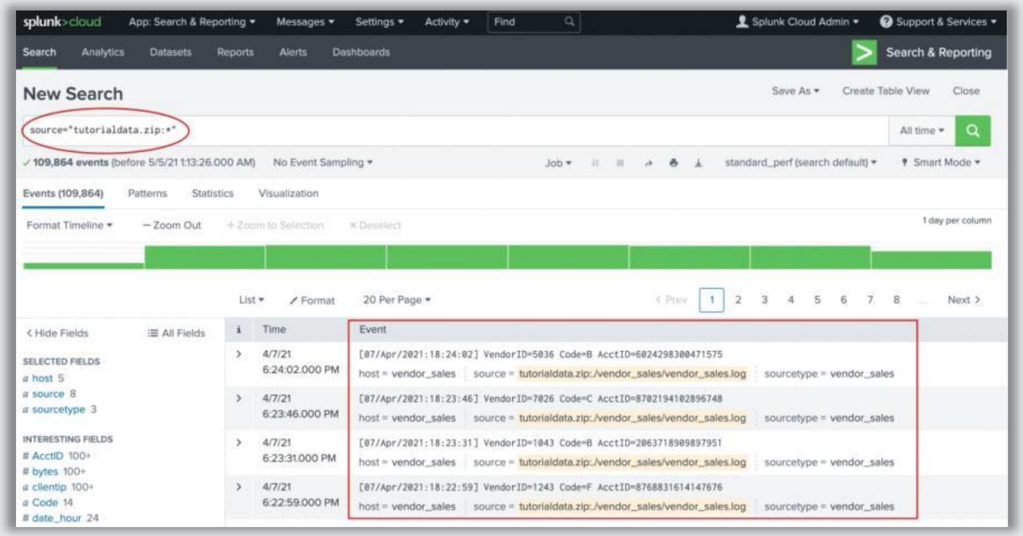
Visualize your searches. Learn more.

# Splunk cloud

- To see the data in the Search app, click Start Searching.
- You might see a screen asking if you want to take a tour. You can take the tour or click Skip. The Search app opens and a search is automatically run on the tutorial data source.

# Splunk cloud

- To see the data in the Search app, click Start Searching.
- You might see a screen asking if you want to take a tour. You can take the tour or click Skip.
- The Search app opens and a search is automatically run on the tutorial data source



New Search

```
source="tutorialdata.zip:*"
```

✓ 109,864 events (before 5/5/21 1:13:26.000 AM)

Events (109,864)

SELECTED FIELDS
a host 5
a source 8
a sourcetype 3

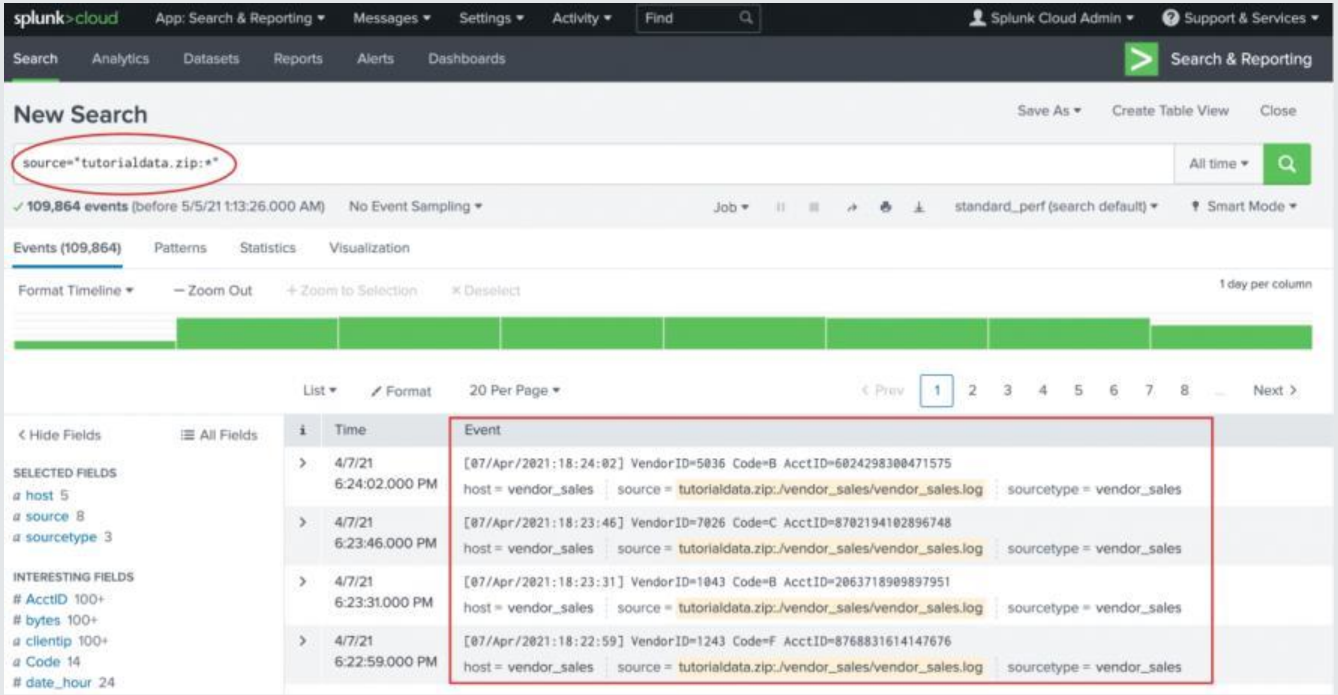
INTERESTING FIELDS
# AcctID 100+
# bytes 100+
a clientip 100+
a Code 14
# date_hour 24

| i | Time | Event |
|---|---|---|
| › | 4/7/21<br>6:24:02.000 PM | [07/Apr/2021:18:24:02] VendorID=5036 Code=B AcctID=6024298300471575<br>host = vendor_sales source = tutorialdata.zip:./vendor_sales/vendor_sales.log sourcetype = vendor_sales |
| › | 4/7/21<br>6:23:46.000 PM | [07/Apr/2021:18:23:46] VendorID=7026 Code=C AcctID=8702194102896748<br>host = vendor_sales source = tutorialdata.zip:./vendor_sales/vendor_sales.log sourcetype = vendor_sales |
| › | 4/7/21<br>6:23:31.000 PM | [07/Apr/2021:18:23:31] VendorID=1043 Code=B AcctID=2063718909897951<br>host = vendor_sales source = tutorialdata.zip:./vendor_sales/vendor_sales.log sourcetype = vendor_sales |
| › | 4/7/21<br>6:22:59.000 PM | [07/Apr/2021:18:22:59] VendorID=1243 Code=F AcctID=8768831614147676<br>host = vendor_sales source = tutorialdata.zip:./vendor_sales/vendor_sales.log sourcetype = vendor_sales |

# Splunk cloud

Success! The results confirm that the data in zip file was indexed and that events were created.

Click the Splunk logo to return to Splunk Home.



New Search

```
source="tutorialdata.zip:*"
```

✓ 109,864 events (before 5/5/21 1:13:26.000 AM)   No Event Sampling ▾   standard_perf (search default) ▾

Events (109,864)

SELECTED FIELDS
- host 5
- source 8
- sourcetype 3

INTERESTING FIELDS
- AcctID 100+
- bytes 100+
- clientip 100+
- Code 14
- date_hour 24

```
i  Time                  Event
>  4/7/21                [07/Apr/2021:18:24:02] VendorID=5036 Code=B AcctID=6024298300471575
   6:24:02.000 PM        host = vendor_sales   source = tutorialdata.zip:./vendor_sales/vendor_sales.log   sourcetype = vendor_sales
>  4/7/21                [07/Apr/2021:18:23:46] VendorID=7026 Code=C AcctID=8702194102896748
   6:23:46.000 PM        host = vendor_sales   source = tutorialdata.zip:./vendor_sales/vendor_sales.log   sourcetype = vendor_sales
>  4/7/21                [07/Apr/2021:18:23:31] VendorID=1043 Code=B AcctID=2063718909897951
   6:23:31.000 PM        host = vendor_sales   source = tutorialdata.zip:./vendor_sales/vendor_sales.log   sourcetype = vendor_sales
>  4/7/21                [07/Apr/2021:18:22:59] VendorID=1243 Code=F AcctID=8768831614147676
   6:22:59.000 PM        host = vendor_sales   source = tutorialdata.zip:./vendor_sales/vendor_sales.log   sourcetype = vendor_sales
```

That’s all for now...