

# 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?

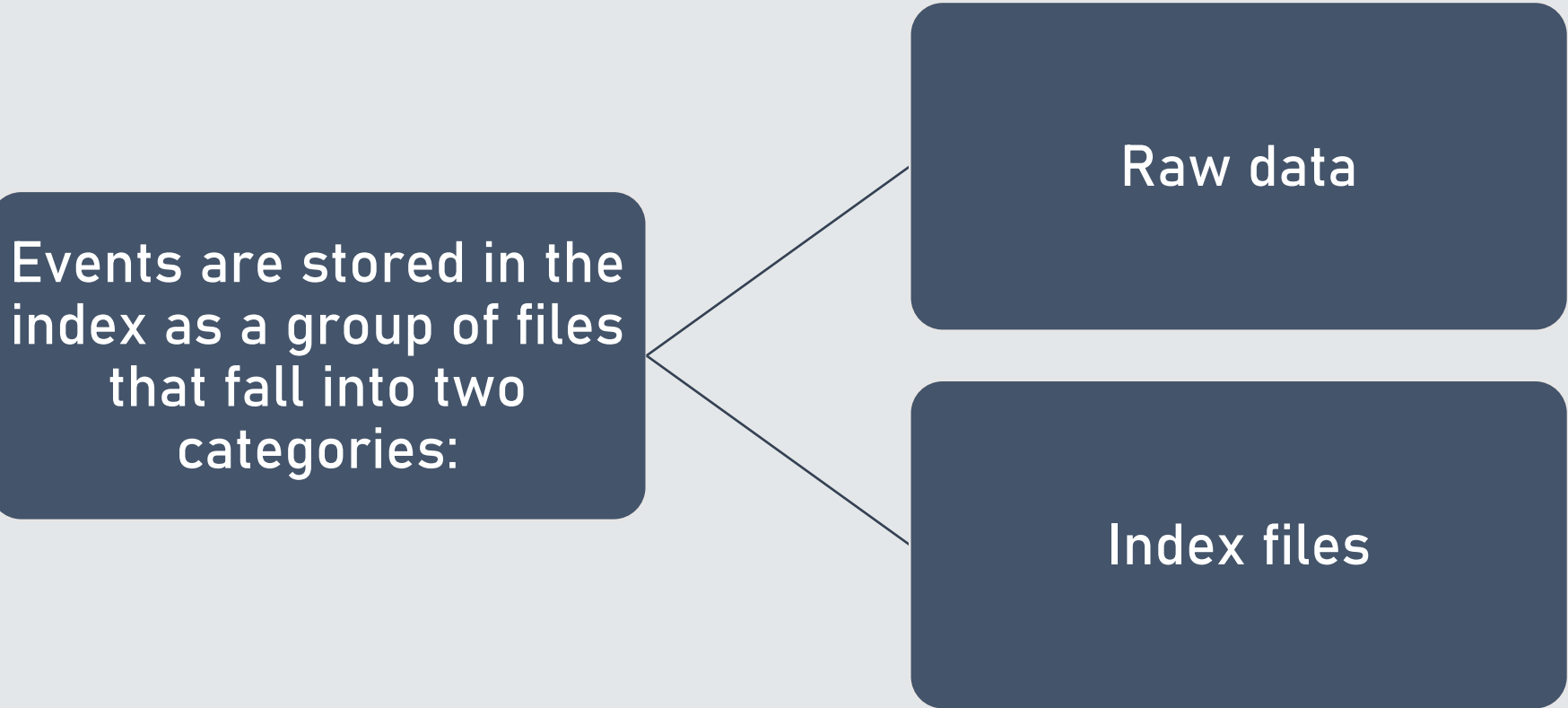
Data source
Files and directories
Network events
IT Operations
Cloud services
Database services
Security services
Virtualization services
Application servers
Windows sources
Other sources

# 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:



```
graph LR; A[Events are stored in the index as a group of files that fall into two categories:] --> B[Raw data]; A --> C[Index files];
```

Raw data

Index files

# Where is the data stored?

Buckets

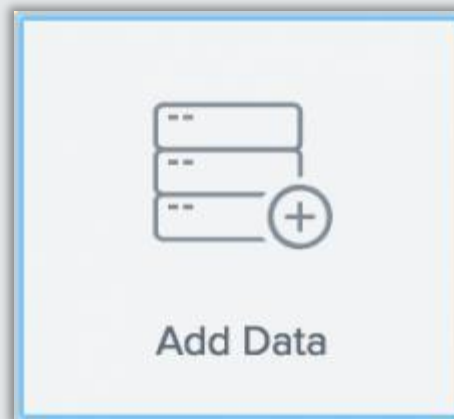
Main

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.



# Use the Add Data wizard

Under Select Source, click Select File..

The screenshot shows the 'Add Data' wizard in Splunk. At the top, a progress bar indicates the current step is 'Select Source', followed by 'Set Source Type', 'Input Settings', 'Review', and 'Done'. Navigation buttons for '< Back' and 'Next >' are on the right. The main section is titled 'Select Source' and contains the instruction: '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](#)'. Below this, it says 'Selected File: No file selected'. A button labeled 'Select File' is circled in red. Underneath is a large rectangular drop zone with the text 'Drop your data file here' and a note: 'The maximum file upload size is 500 Mb'. At the bottom, there is an 'FAQ' section with three questions: 'What kinds of files can the Splunk platform index?', 'What is a source?', and 'How do I get remote data onto my Splunk platform instance?'.

**Add Data**

Select Source Set Source Type Input Settings Review Done

< Back Next >

### 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

Select File

Drop your data file here

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

Add Data

< Back

Review >

## 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.

AutomaticSelectNew

### 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

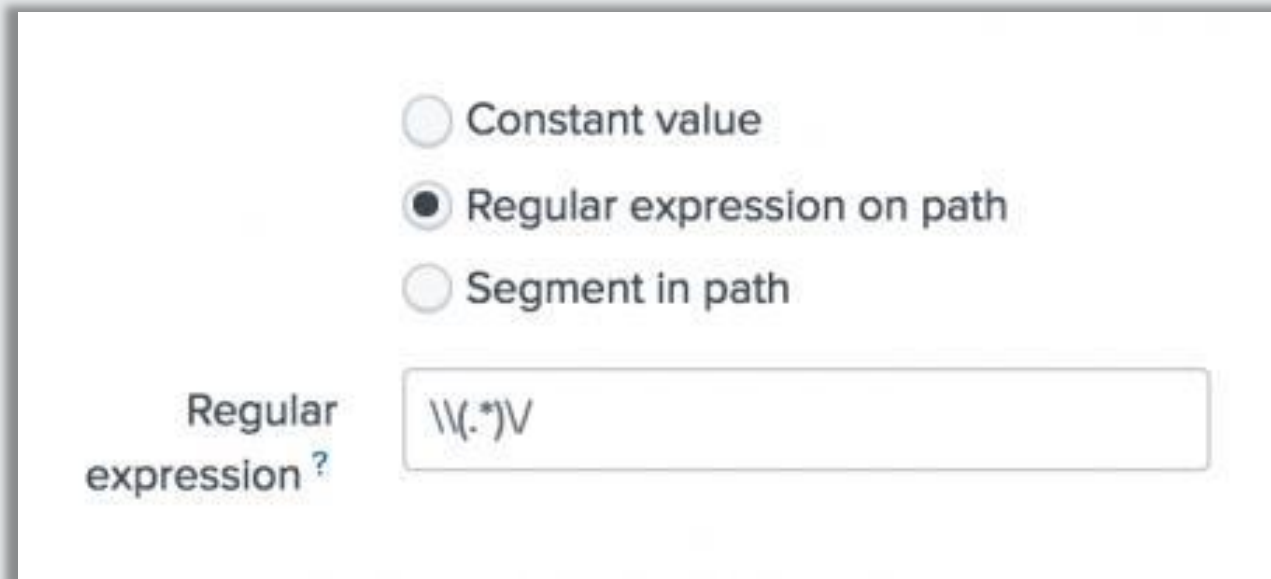
### Index

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](#)

IndexDefaultCreate a new index

# Use the Add Data wizard

- a. Select Regular expression on path.
- b. Type `\\(.*)\\` for the regex to extract the host values from the path.



The screenshot shows a dialog box with three radio button options. The second option, 'Regular expression on path', is selected. Below the options is a text input field containing the regular expression '\\(.\*)\\'. To the left of the input field is the label 'Regular expression ?'.

☐ Constant value

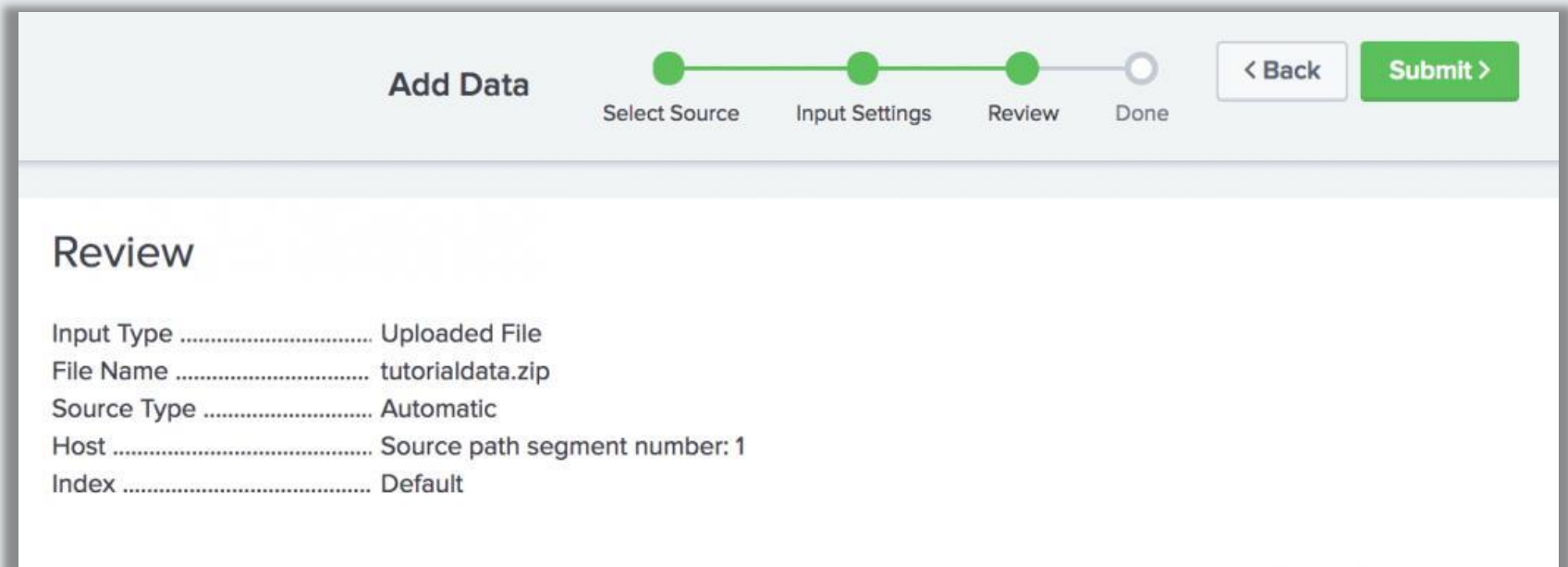
☒ Regular expression on path

☐ Segment in path

Regular expression ?

# Use the Add Data wizard

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



The screenshot shows the 'Add Data' wizard interface. At the top, there is a progress bar with four steps: 'Select Source', 'Input Settings', 'Review', and 'Done'. The 'Review' step is currently active, indicated by a green dot and a green line connecting the first three steps. To the right of the progress bar are two buttons: '< Back' and 'Submit >'. Below the progress bar, the title 'Review' is displayed. Underneath the title, the following input settings are listed:

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.

Add Data

✓

Select SourceInput SettingsReviewDone

< BackNext >

✓

File has been uploaded successfully.

Configure your inputs by going to Settings > [Data Inputs](#)

Start Searching

Search your data now or see [examples and tutorials](#). [🔗](#)

Add More Data

Add more data inputs now or see [examples and tutorials](#). [🔗](#)

Download Apps

Apps help you do more with your data. [Learn more](#). [🔗](#)

Build Dashboards

Visualize your searches. [Learn more](#). [🔗](#)

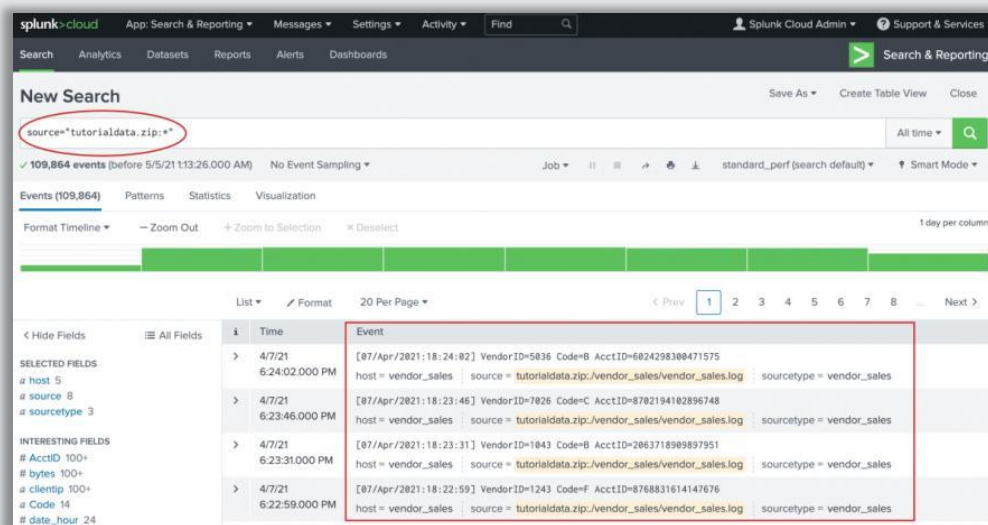


# Splunk cloud

- 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



# 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.

The screenshot displays the Splunk Cloud Search & Reporting interface. At the top, the navigation bar includes the Splunk logo, user profile (Splunk Cloud Admin), and support links. Below this, a secondary navigation bar shows various app categories like Search, Analytics, and Reports. The main section is titled 'New Search' and features a search bar with the query `source=*tutorialdata.zip:*`, which is circled in red. Below the search bar, a status bar indicates '109,864 events' and provides options for saving, creating table views, and closing the search. The search results are displayed in a table format, with a timeline visualization at the top. The table has columns for 'Time' and 'Event'. The 'Event' column contains detailed log entries, including timestamps, VendorID, Code, AcctID, host, source, and sourcetype. A red box highlights the first four rows of the table. The interface also includes a sidebar with 'SELECTED FIELDS' and 'INTERESTING FIELDS' on the left, and a bottom section with pagination and format options.

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



**That's all for now...**