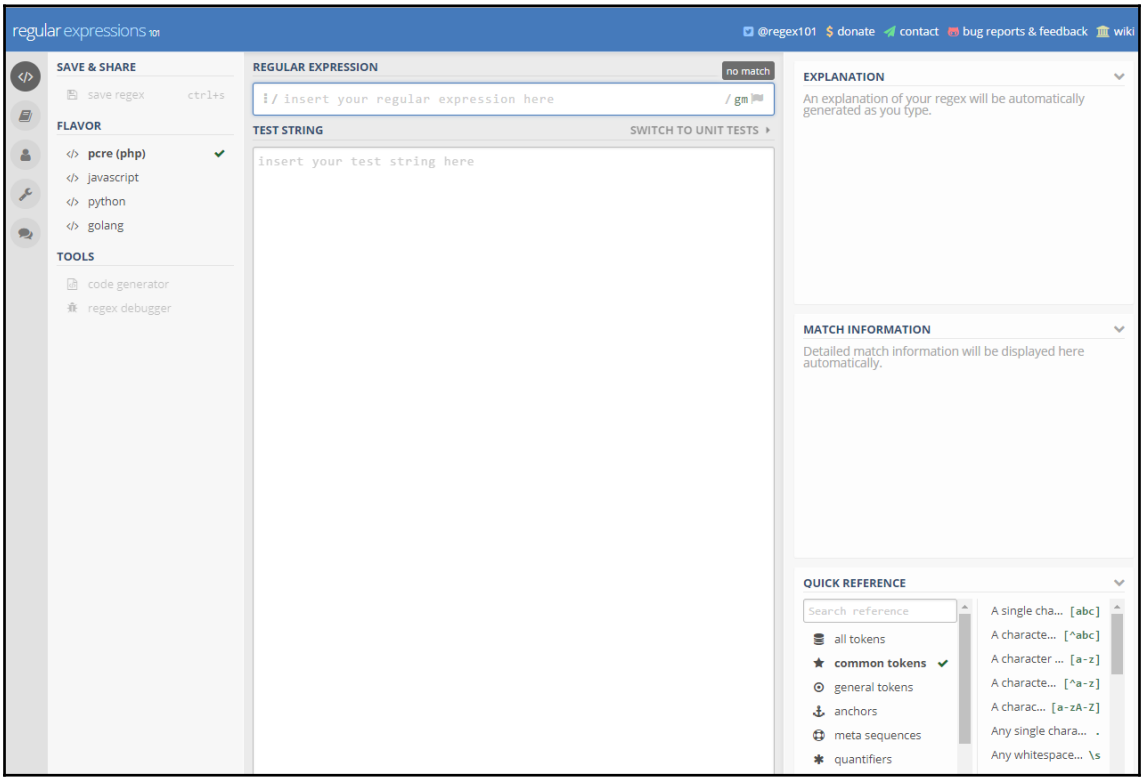
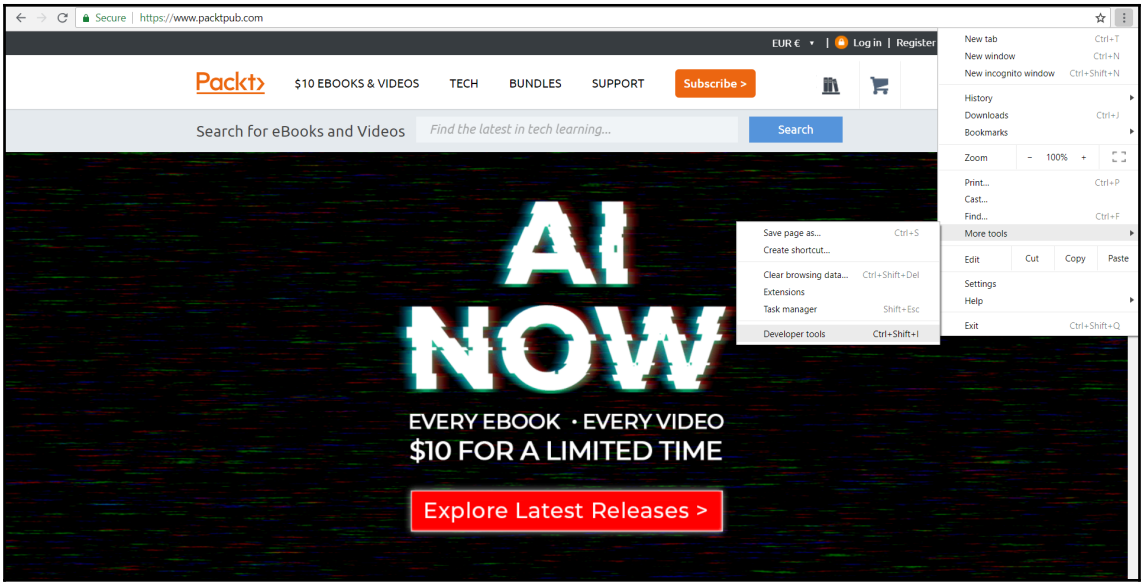
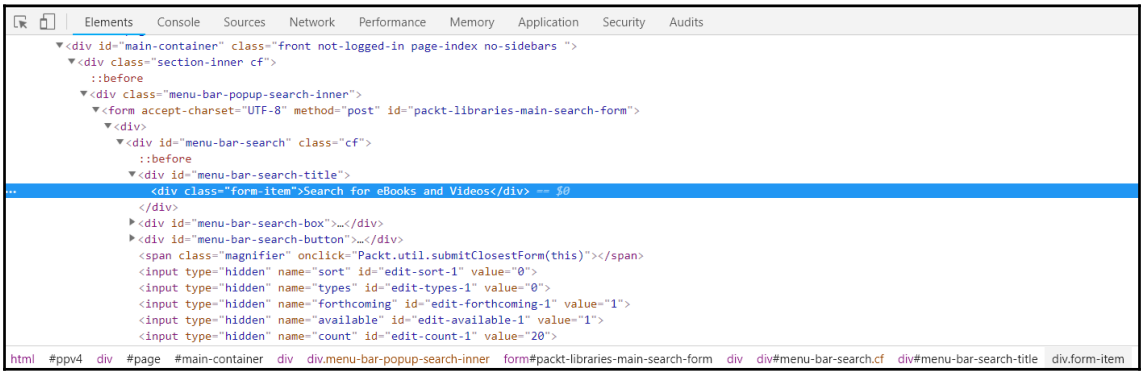
Chapter 1: Introduction to Web Scraping

Chapter 2:
XML Path Language and Regular Expression Language




Secure | https://www.packtpub.com
EUR € | Log in | Register
Packt
$10 EBOOKS & VIDEOS
TECH
BUNDLES
SUPPORT
Subscribe >
Search for eBooks and Videos
Find the latest in tech learning...
Search
AI NOW
EVERY EBOOK · EVERY VIDEO
$10 FOR A LIMITED TIME
Explore Latest Releases >
New tab Ctrl+T
New window Ctrl+N
New incognito window Ctrl+Shift+N
History
Downloads Ctrl+J
Bookmarks
Zoom − 100% +
Print... Ctrl+P
Cast...
Find... Ctrl+F
More tools
Edit Cut Copy Paste
Settings
Help
Exit Ctrl+Shift+Q
Save page as... Ctrl+S
Create shortcut...
Clear browsing data... Ctrl+Shift+Del
Extensions
Task manager Shift+Esc
Developer tools Ctrl+Shift+I


Elements Console Sources Network Performance Memory Application Security Audits

::before

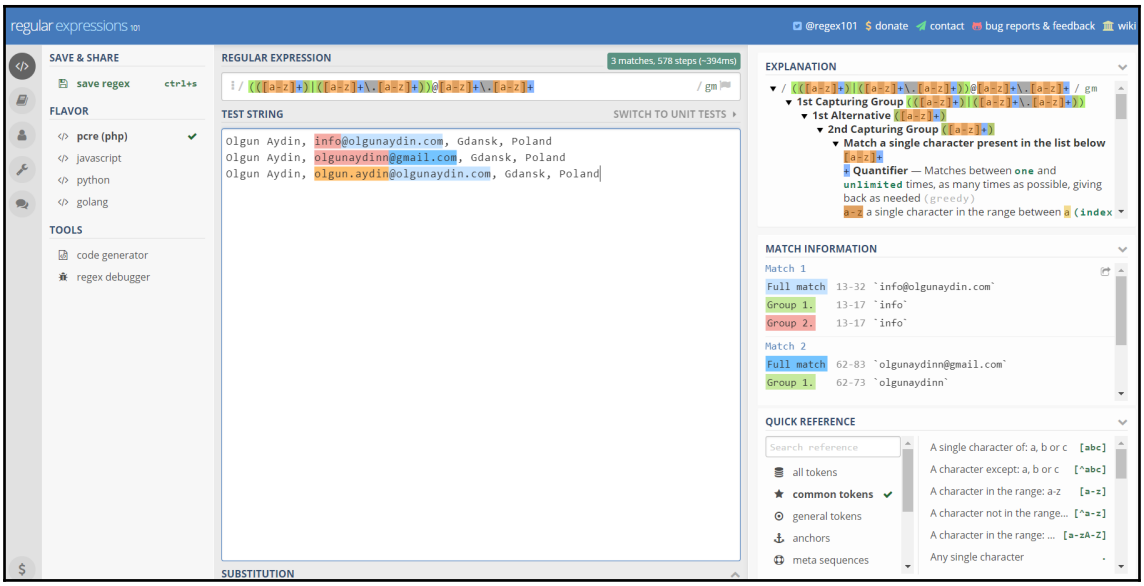
<form accept-charset="UTF-8" method="post" id="packt-libraries-main-search-form">

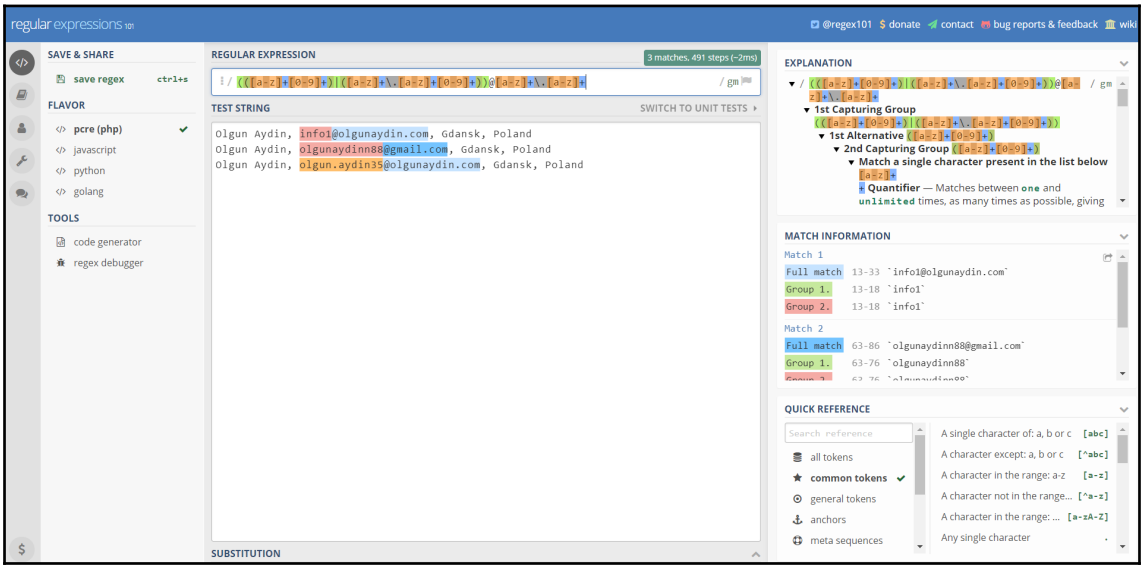
::before

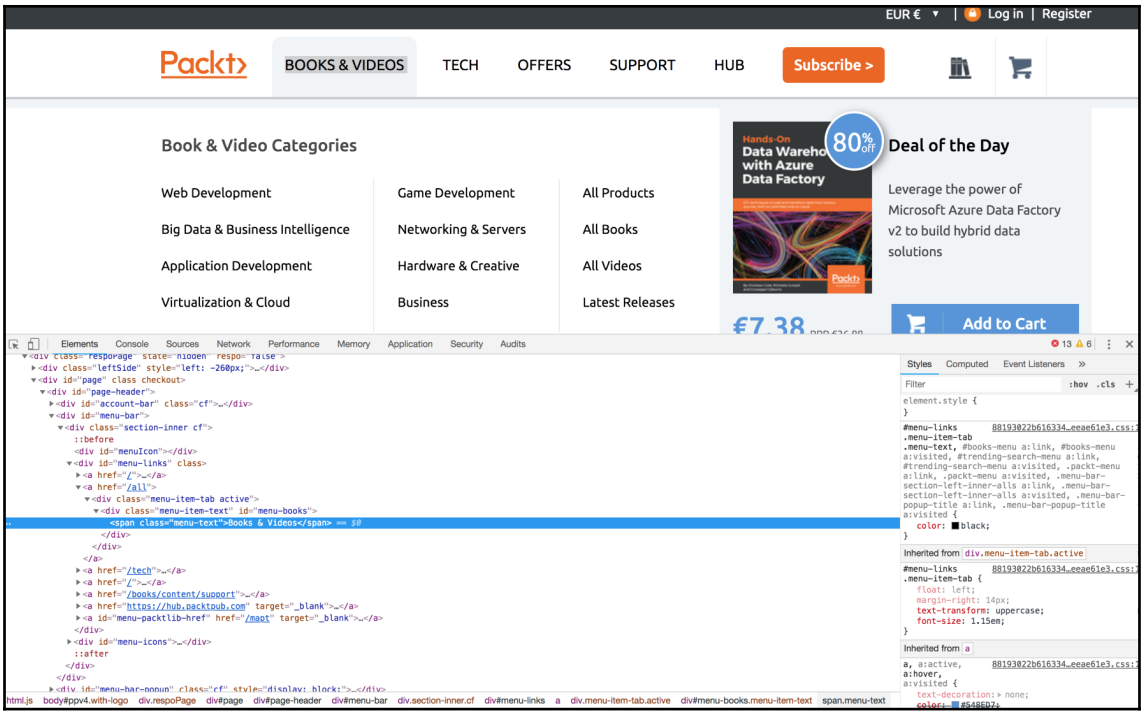
Search for eBooks and Videos == $0

…
…

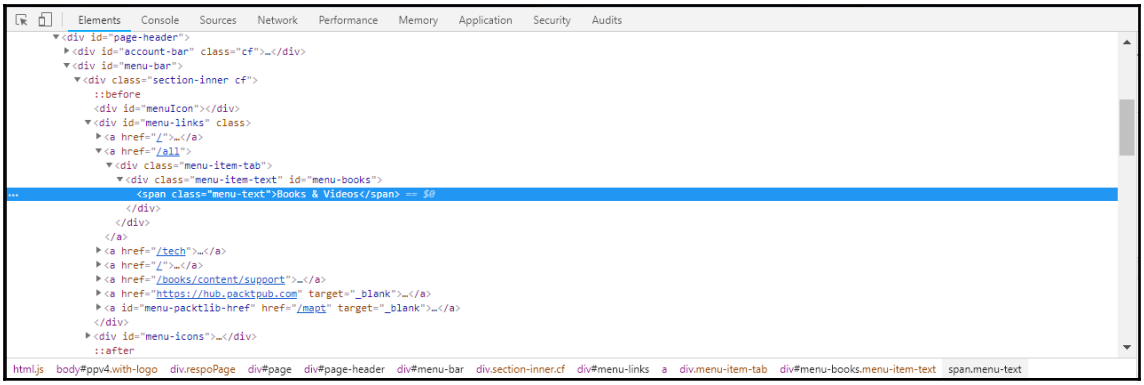
<input type="hidden" name="sort" id="edit-sort-1" value="0">
<input type="hidden" name="types" id="edit-types-1" value="0">
<input type="hidden" name="forthcoming" id="edit-forthcoming-1" value="1">
<input type="hidden" name="available" id="edit-available-1" value="1">
<input type="hidden" name="count" id="edit-count-1" value="20">
html #ppv4 div #page #main-container div div.menu-bar-popup-search-inner form#packt-libraries-main-search-form div div#menu-bar-search.cf div#menu-bar-search-title div.form-item

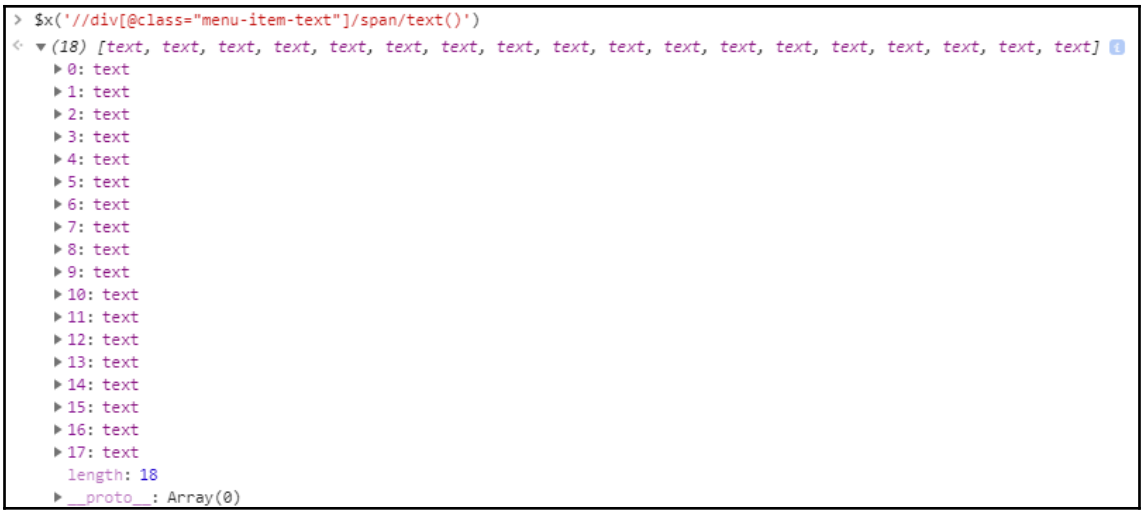

regular expressions 101
@regex101 $ donate contact bug reports & feedback wiki
SAVE & SHARE
save regex ctrl+s
FLAVOR
pcre (php)
javascript
python
golang
TOOLS
code generator
regex debugger
REGULAR EXPRESSION
3 matches, 578 steps (~394ms)
/ (([a-z]+)|([a-z]+\.[a-z]+))@[a-z]+\.[a-z]+ / gm
TEST STRING
SWITCH TO UNIT TESTS
Olgun Aydin, info@olgunaydin.com, Gdansk, Poland
Olgun Aydin, olgunaydinn@gmail.com, Gdansk, Poland
Olgun Aydin, olgun.aydin@olgunaydin.com, Gdansk, Poland
SUBSTITUTION
EXPLANATION
1st Capturing Group
1st Alternative
2nd Capturing Group
Match a single character present in the list below
Quantifier — Matches between one and unlimited times, as many times as possible, giving back as needed (greedy)
a-z a single character in the range between a (index
MATCH INFORMATION
Match 1
Full match 13-32 `info@olgunaydin.com`
Group 1. 13-17 `info`
Group 2. 13-17 `info`
Match 2
Full match 62-83 `olgunaydinn@gmail.com`
Group 1. 62-73 `olgunaydinn`
QUICK REFERENCE
Search reference
all tokens
common tokens
general tokens
anchors
meta sequences
A single character of: a, b or c [abc]
A character except: a, b or c [^abc]
A character in the range: a-z [a-z]
A character not in the range... [^a-z]
A character in the range: ... [a-zA-Z]
Any single character .

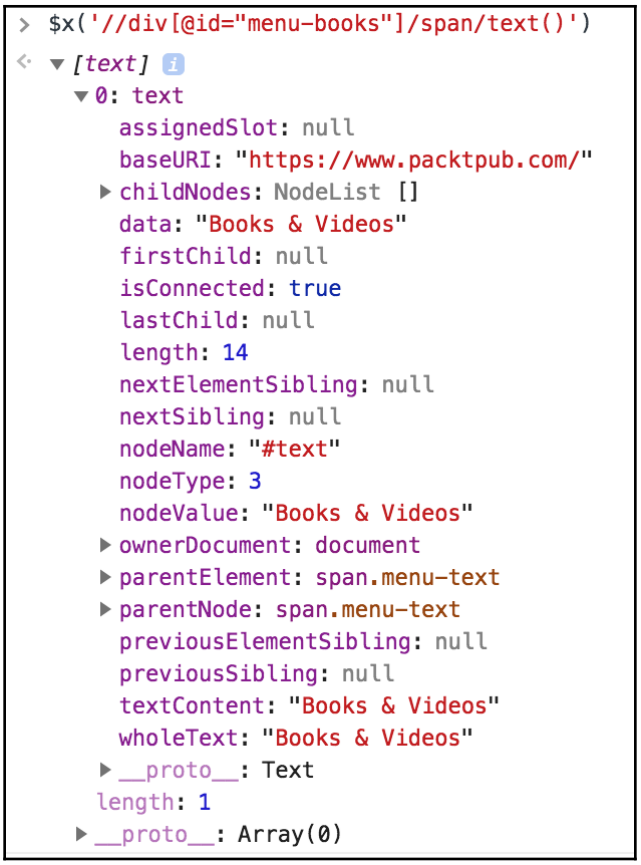

regular expressions 101
@regex101 $ donate contact bug reports & feedback wiki
SAVE & SHARE
save regex ctrl+s
FLAVOR
pcre (php)
javascript
python
golang
TOOLS
code generator
regex debugger
REGULAR EXPRESSION
3 matches, 491 steps (~2ms)
/ (([a-z]+[0-9]+)|([a-z]+\.[a-z]+[0-9]+))@[a-z]+\.[a-z]+ / gm
TEST STRING
SWITCH TO UNIT TESTS
Olgun Aydin, info1@olgunaydin.com, Gdansk, Poland
Olgun Aydin, olgunaydinn88@gmail.com, Gdansk, Poland
Olgun Aydin, olgun.aydin35@olgunaydin.com, Gdansk, Poland
SUBSTITUTION
EXPLANATION
1st Capturing Group
1st Alternative
2nd Capturing Group
Match a single character present in the list below
Quantifier — Matches between one and unlimited times, as many times as possible, giving
MATCH INFORMATION
Match 1
Full match 13-33 `info1@olgunaydin.com`
Group 1. 13-18 `info1`
Group 2. 13-18 `info1`
Match 2
Full match 63-86 `olgunaydinn88@gmail.com`
Group 1. 63-76 `olgunaydinn88`
QUICK REFERENCE
Search reference
all tokens
common tokens
general tokens
anchors
meta sequences
A single character of: a, b or c [abc]
A character except: a, b or c [^abc]
A character in the range: a-z [a-z]
A character not in the range... [^a-z]
A character in the range: ... [a-zA-Z]
Any single character .


EUR €
Log in | Register
Packt
BOOKS & VIDEOS
TECH
OFFERS
SUPPORT
HUB
Subscribe >
Book & Video Categories
Web Development
Big Data & Business Intelligence
Application Development
Virtualization & Cloud
Game Development
Networking & Servers
Hardware & Creative
Business
All Products
All Books
All Videos
Latest Releases
Hands-On
Data Warehouse with Azure Data Factory
80% off
Deal of the Day
Leverage the power of Microsoft Azure Data Factory v2 to build hybrid data solutions
Add to Cart
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
Styles
Computed
Event Listeners
Filter
element.style {
}
#menu-links .menu-item-tab .menu-text, #books-menu a:link, #books-menu a:visited, #trending-search-menu a:link, #trending-search-menu a:visited, .packt-menu a:link, .packt-menu a:visited, .menu-bar-section-left-inner-alls a:link, .menu-bar-section-left-inner-alls a:visited, .menu-bar-popup-title a:link, .menu-bar-popup-title a:visited {
color: black;
}
Inherited from div.menu-item-tab.active
#menu-links .menu-item-tab {
float: left;
margin-right: 14px;
text-transform: uppercase;
font-size: 1.15em;
}
Inherited from a
a, a:active, a:hover, a:visited {
text-decoration: none;

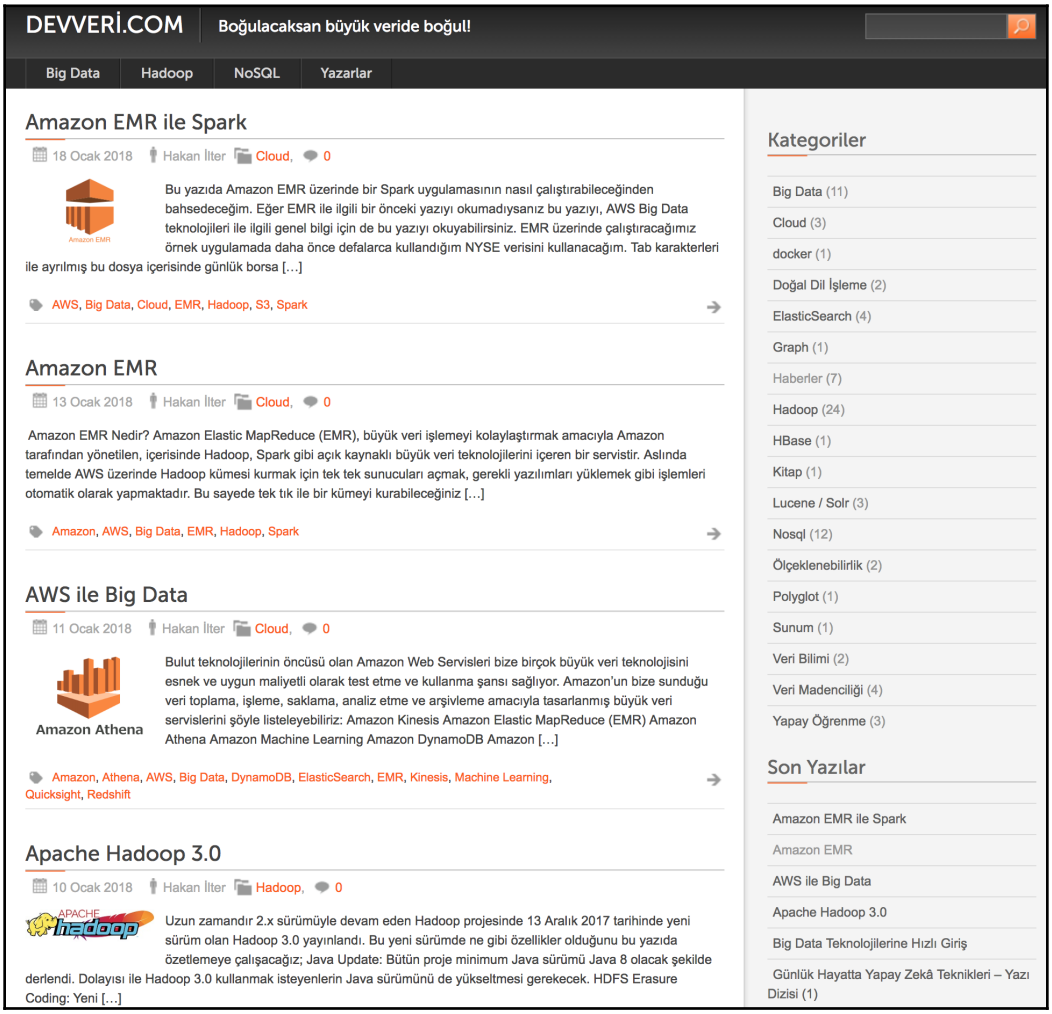
```
> $x('//*[@id="menu-books"]')
▼ [div#menu-books.menu-item-text]
  ▼ 0: div#menu-books.menu-item-text
      accessKey: ""
      align: ""
      assignedSlot: null
    ▶ attributeStyleMap: StylePropertyMap {size: 0}
    ▶ attributes: NamedNodeMap {0: class, 1: id, class: class, id: id, length: 2}
      autocapitalize: ""
      baseURI: "https://www.packtpub.com/"
      childElementCount: 1
    ▶ childNodes: NodeList(3) [text, span.menu-text, text]
    ▶ children: HTMLCollection [span.menu-text]
    ▶ classList: DOMTokenList ["menu-item-text", value: "menu-item-text"]
      className: "menu-item-text"
      clientHeight: 65
      clientLeft: 0
      clientTop: 0
      clientWidth: 156
      contentEditable: "inherit"
    ▶ dataset: DOMStringMap {}
      dir: ""
      draggable: false
    ▶ firstChild: text
    ▶ firstElementChild: span.menu-text
      hidden: false
```



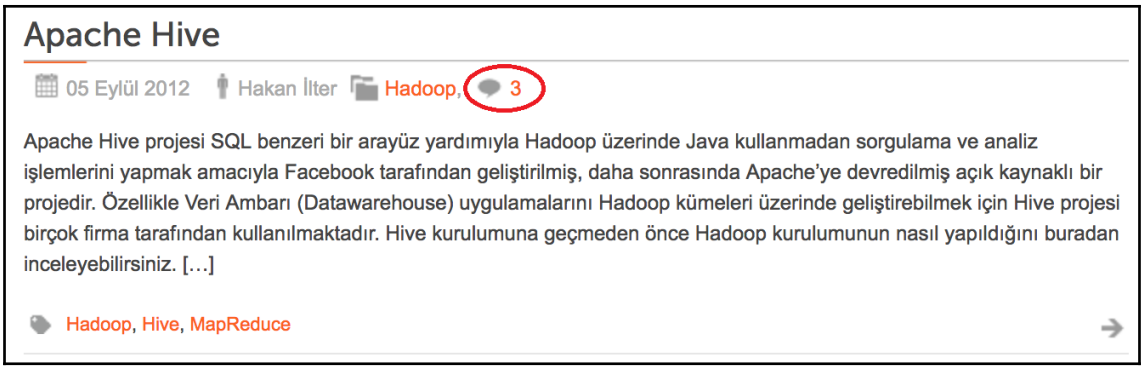

$x('//div[@class="menu-item-text"]/span/text()')
(18) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]
0: text
1: text
2: text
3: text
4: text
5: text
6: text
7: text
8: text
9: text
10: text
11: text
12: text
13: text
14: text
15: text
16: text
17: text
length: 18
__proto__: Array(0)

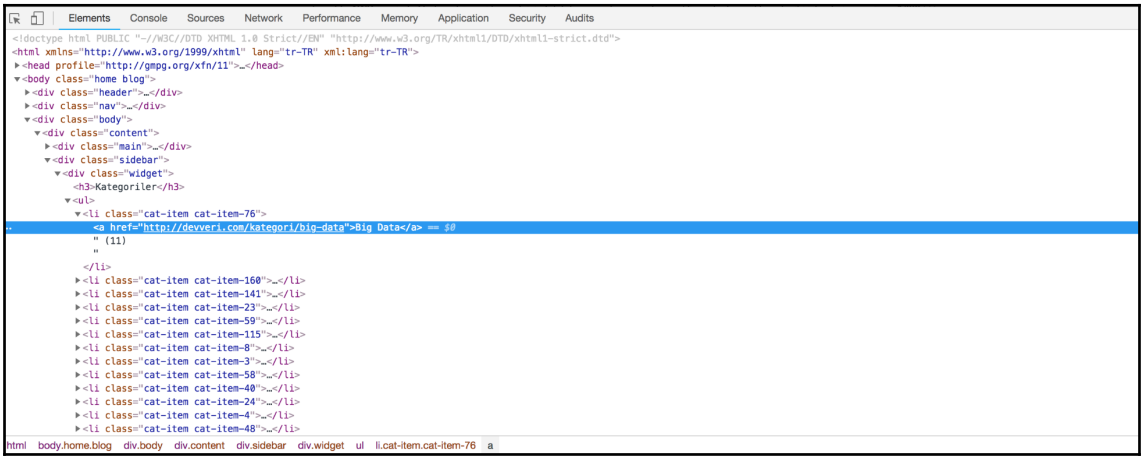

$x('//div[@id="menu-books"]/span/text()')
[text]
0: text
assignedSlot: null
baseURI: "https://www.packtpub.com/"
childNodes: NodeList []
data: "Books & Videos"
firstChild: null
isConnected: true
lastChild: null
length: 14
nextElementSibling: null
nextSibling: null
nodeName: "#text"
nodeType: 3
nodeValue: "Books & Videos"
ownerDocument: document
parentElement: span.menu-text
parentNode: span.menu-text
previousElementSibling: null
previousSibling: null
textContent: "Books & Videos"
wholeText: "Books & Videos"
__proto__: Text
length: 1
__proto__: Array(0)

Chapter 3: Web Scraping with rvest




```
> urls
[1] "http://devveri.com/"
```


Apache Hive
05 Eylül 2012
Hakan İlter
Hadoop,
3
Apache Hive projesi SQL benzeri bir arayüz yardımıyla Hadoop üzerinde Java kullanmadan sorgulama ve analiz işlemlerini yapmak amacıyla Facebook tarafından geliştirilmiş, daha sonrasında Apache'ye devredilmiş açık kaynaklı bir projedir. Özellikle Veri Ambarı (Datawarehouse) uygulamalarını Hadoop kümeleri üzerinde geliştirebilmek için Hive projesi birçok firma tarafından kullanılmaktadır. Hive kurulumuna geçmeden önce Hadoop kurulumunun nasıl yapıldığını buradan inceleyebilirsiniz. […]
Hadoop, Hive, MapReduce

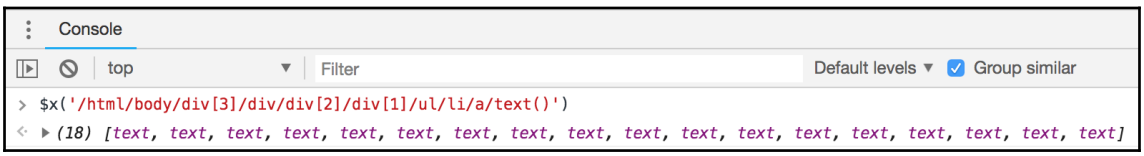

Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
<!doctype html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="tr-TR" xml:lang="tr-TR">
<head profile="http://gmpg.org/xfn/11">…</head>
<body class="home blog">
…
…

…

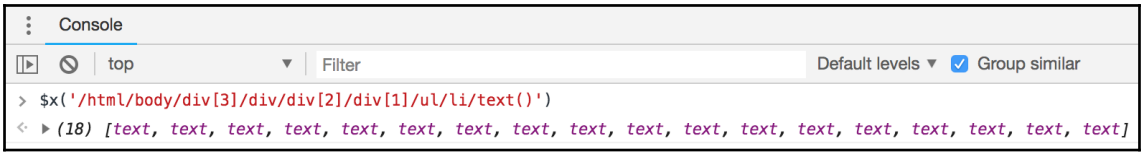
<h3>Kategoriler</h3>

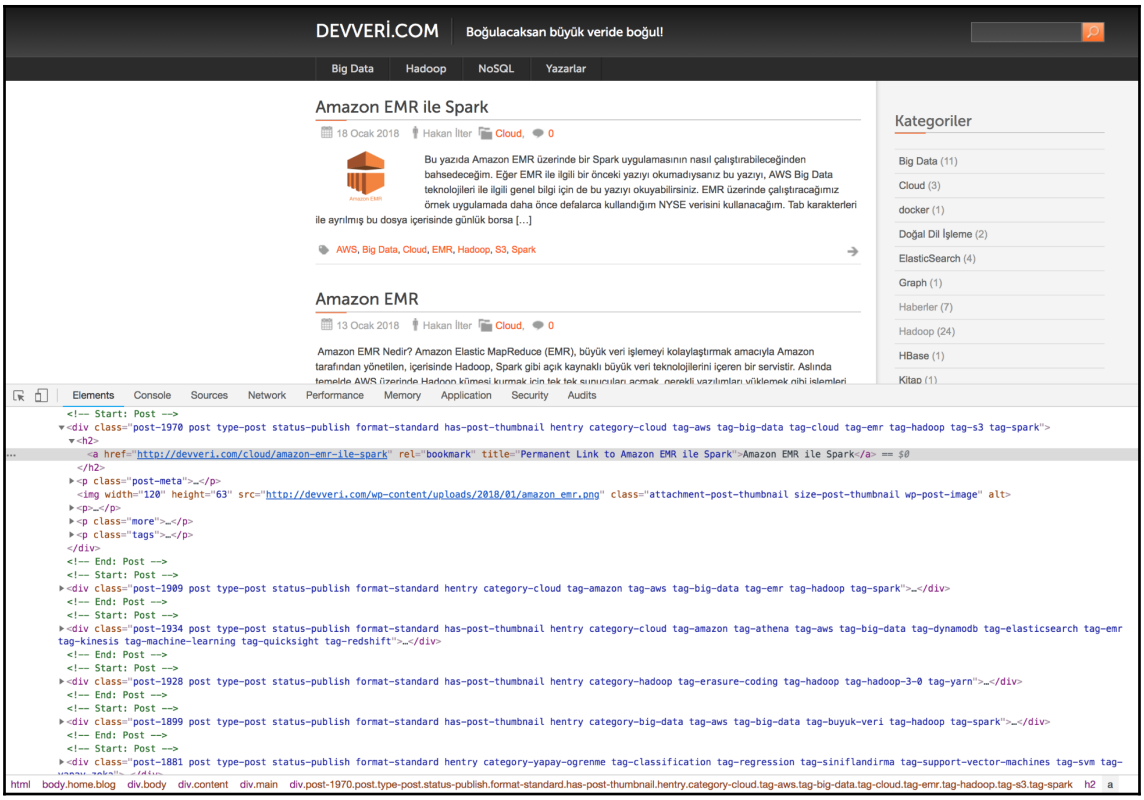
<li class="cat-item cat-item-76">
Big Data == $0
" (11)
"

<li class="cat-item cat-item-160">…
<li class="cat-item cat-item-141">…
<li class="cat-item cat-item-23">…
<li class="cat-item cat-item-59">…
<li class="cat-item cat-item-115">…
<li class="cat-item cat-item-8">…
<li class="cat-item cat-item-3">…
<li class="cat-item cat-item-58">…
<li class="cat-item cat-item-40">…
<li class="cat-item cat-item-24">…
<li class="cat-item cat-item-4">…
<li class="cat-item cat-item-48">…
html
body.home.blog
div.body
div.content
div.sidebar
div.widget
ul
li.cat-item.cat-item-76
a


Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[2]/div[1]/ul/li/a/text()')
(18) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]

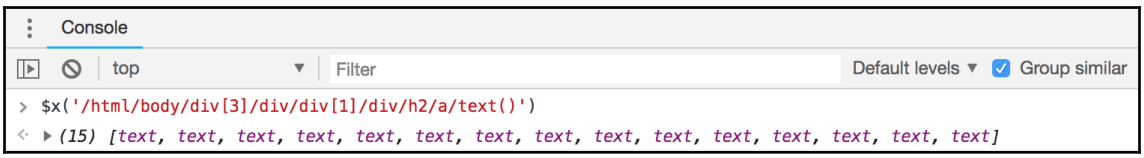
```
> $x('/html/body/div[3]/div/div[2]/div[1]/ul/li[2]/a/text()')
▼ [text] i
  ▼ 0: text
      assignedSlot: null
      baseURI: "http://devveri.com/"
    ▶ childNodes: NodeList []
      data: "Cloud"
      firstChild: null
      isConnected: true
      lastChild: null
      length: 5
      nextElementSibling: null
      nextSibling: null
      nodeName: "#text"
      nodeType: 3
      nodeValue: "Cloud"
    ▶ ownerDocument: document
    ▶ parentElement: a
    ▶ parentNode: a
      previousElementSibling: null
      previousSibling: null
      textContent: "Cloud"
      wholeText: "Cloud"
    ▶ __proto__: Text
    length: 1
  ▶ __proto__: Array(0)
```

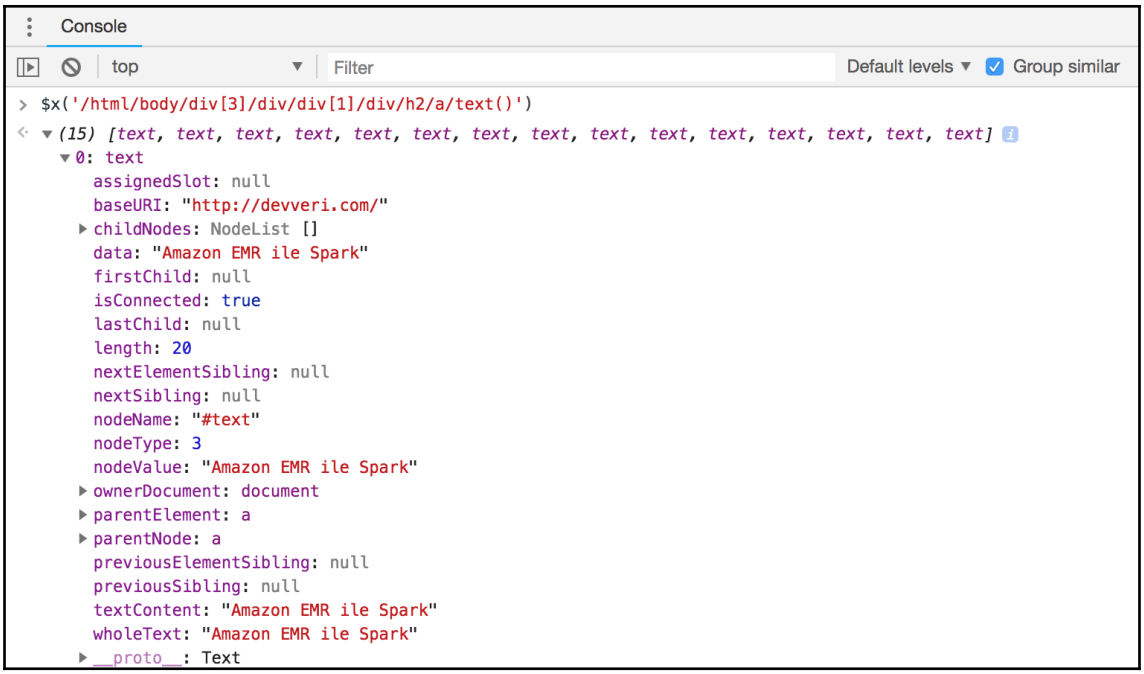


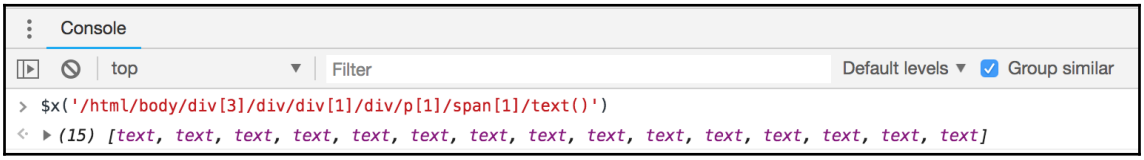

DEVVERİ.COM
Boğulacaksan büyük veride boğul!
Big Data
Hadoop
NoSQL
Yazarlar
Amazon EMR ile Spark
18 Ocak 2018
Hakan İlter
Cloud,
0
Bu yazıda Amazon EMR üzerinde bir Spark uygulamasının nasıl çalıştırabileceğinden bahsedeceğim. Eğer EMR ile ilgili bir önceki yazıyı okumadıysanız bu yazıyı, AWS Big Data teknolojileri ile ilgili genel bilgi için de bu yazıyı okuyabilirsiniz. EMR üzerinde çalıştıracağımız örnek uygulamada daha önce defalarca kullandığım NYSE verisini kullanacağım. Tab karakterleri ile ayrılmış bu dosya içerisinde günlük borsa […]
AWS, Big Data, Cloud, EMR, Hadoop, S3, Spark
Amazon EMR
13 Ocak 2018
Hakan İlter
Cloud,
0
Amazon EMR Nedir? Amazon Elastic MapReduce (EMR), büyük veri işlemeyi kolaylaştırmak amacıyla Amazon tarafından yönetilen, içerisinde Hadoop, Spark gibi açık kaynaklı büyük veri teknolojilerini içeren bir servistir. Aslında
Kategoriler
Big Data (11)
Cloud (3)
docker (1)
Doğal Dil İşleme (2)
ElasticSearch (4)
Graph (1)
Haberler (7)
Hadoop (24)
HBase (1)
Kitap (1)
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
<!-- Start: Post -->

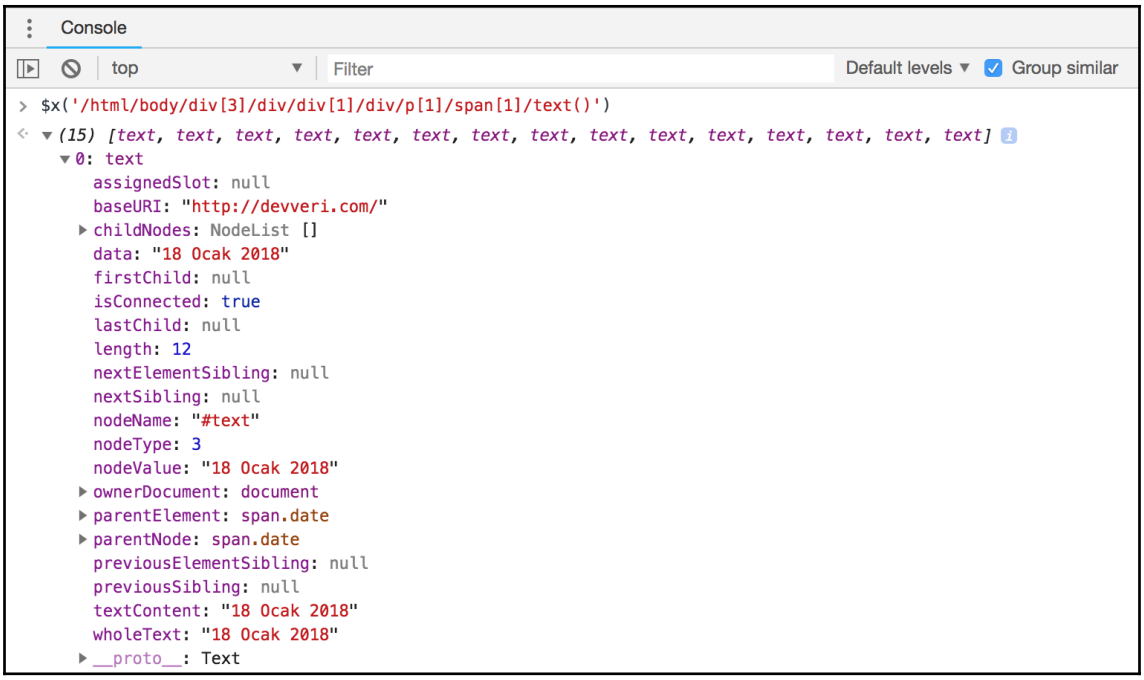
<h2>
Amazon EMR ile Spark == $0
</h2>
…
<img width="120" height="63" src="http://devveri.com/wp-content/uploads/2018/01/amazon_emr.png" class="attachment-post-thumbnail size-post-thumbnail wp-post-image" alt>
…
…
…

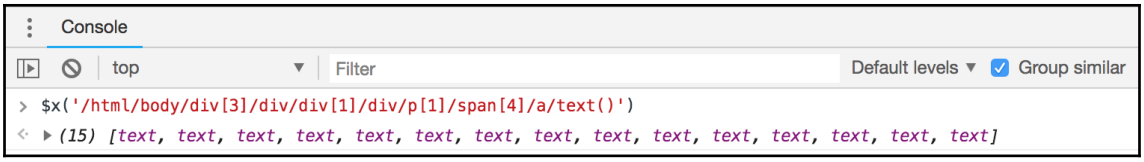
<!-- End: Post -->
<!-- Start: Post -->
…
<!-- End: Post -->
<!-- Start: Post -->
<div class="post-1934 post type-post status-publish format-standard has-post-thumbnail hentry category-cloud tag-amazon tag-athena tag-aws tag-big-data tag-dynamodb tag-elasticsearch tag-emr tag-kinesis tag-machine-learning tag-quicksight tag-redshift">…
<!-- End: Post -->
<!-- Start: Post -->
…
<!-- End: Post -->
<!-- Start: Post -->
…
<!-- End: Post -->
<!-- Start: Post -->
<div class="post-1881 post type-post status-publish format-standard hentry category-yapay-ogrenme tag-classification tag-regression tag-siniflandirma tag-support-vector-machines tag-svm tag-
html
body.home.blog
div.body
div.content
div.main
div.post-1970.post.type-post.status-publish.format-standard.has-post-thumbnail.hentry.category-cloud.tag-aws.tag-big-data.tag-cloud.tag-emr.tag-hadoop.tag-s3.tag-spark
h2
a

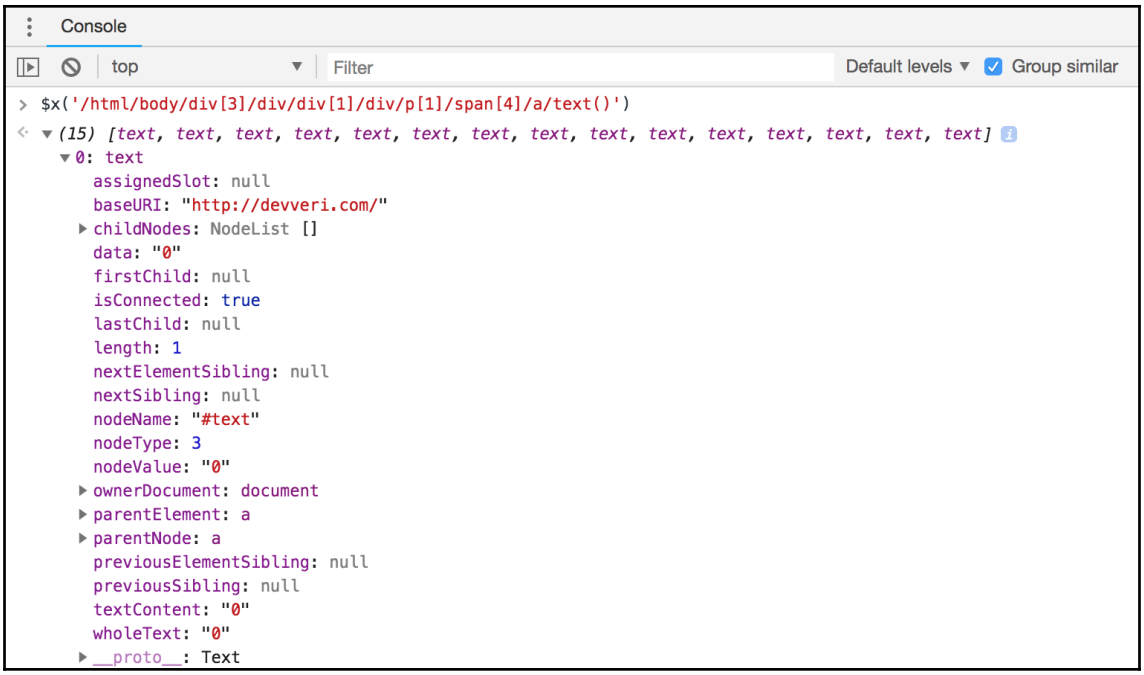

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/h2/a/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]

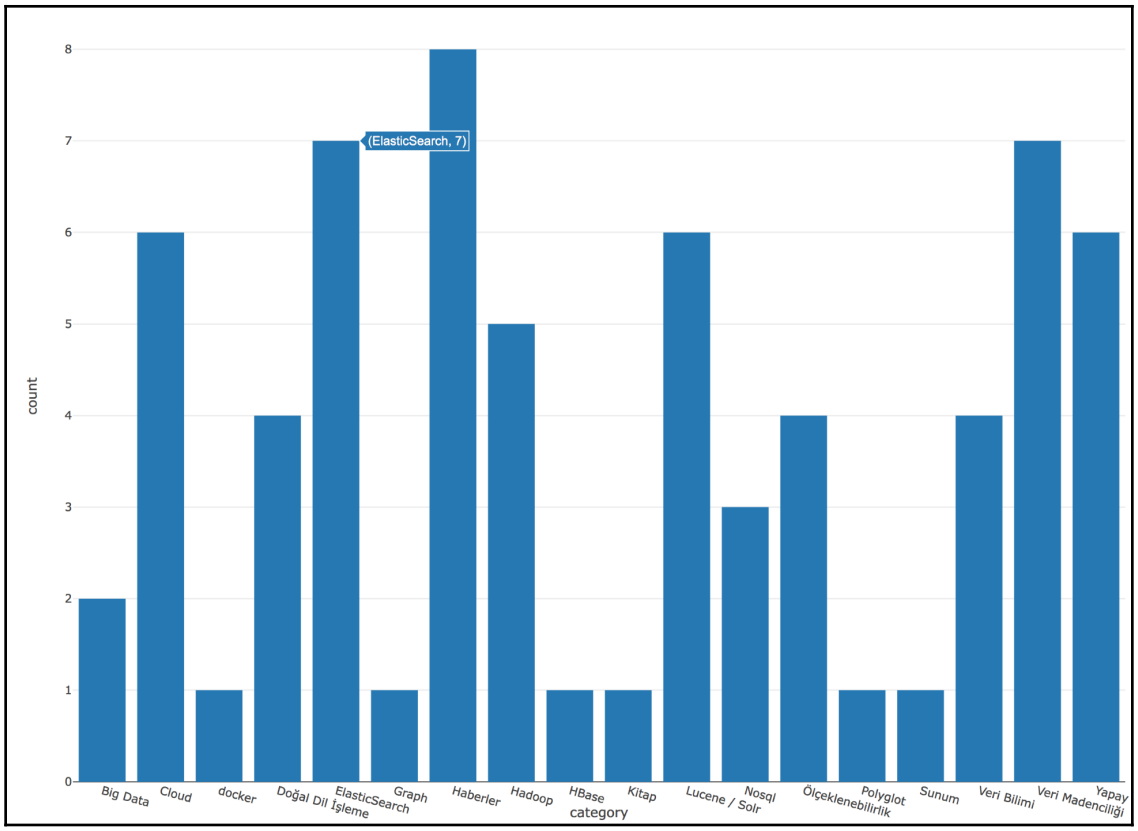

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/h2/a/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]
0: text
assignedSlot: null
baseURI: "http://devveri.com/"
childNodes: NodeList []
data: "Amazon EMR ile Spark"
firstChild: null
isConnected: true
lastChild: null
length: 20
nextElementSibling: null
nextSibling: null
nodeName: "#text"
nodeType: 3
nodeValue: "Amazon EMR ile Spark"
ownerDocument: document
parentElement: a
parentNode: a
previousElementSibling: null
previousSibling: null
textContent: "Amazon EMR ile Spark"
wholeText: "Amazon EMR ile Spark"
__proto__: Text

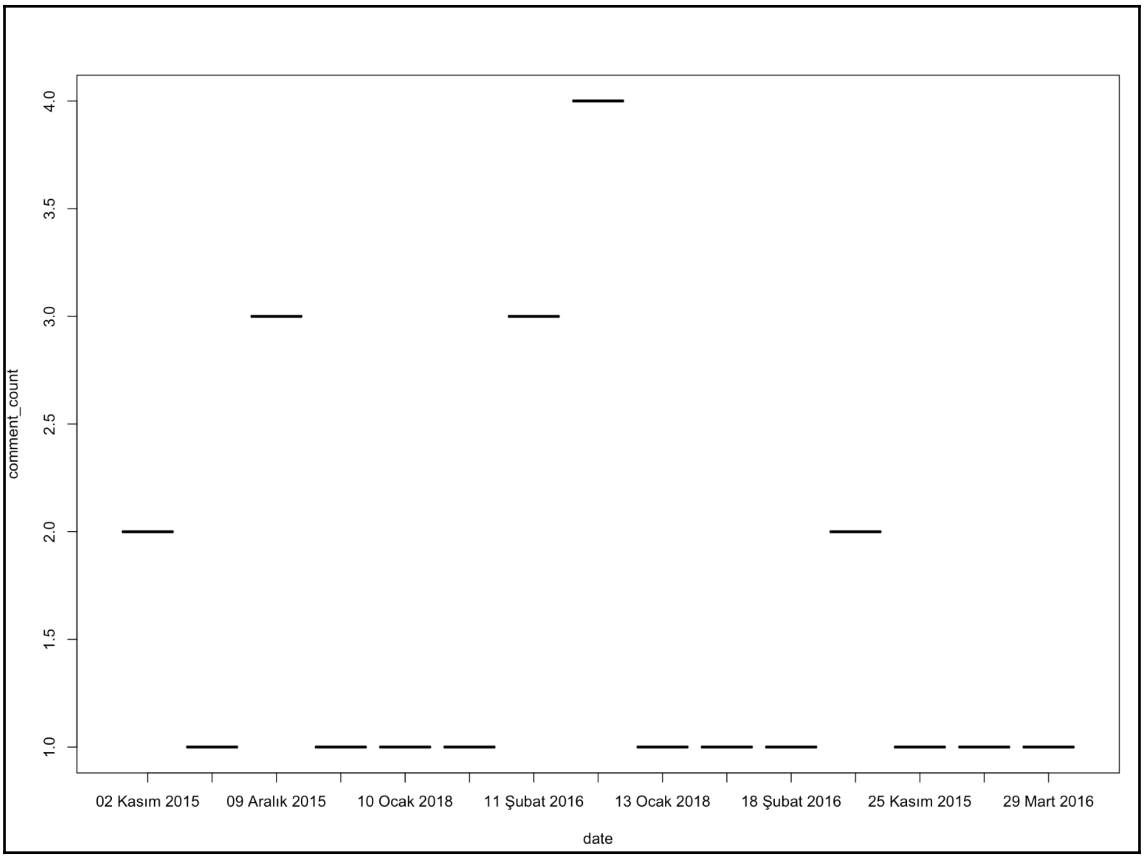

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/p[1]/span[1]/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]

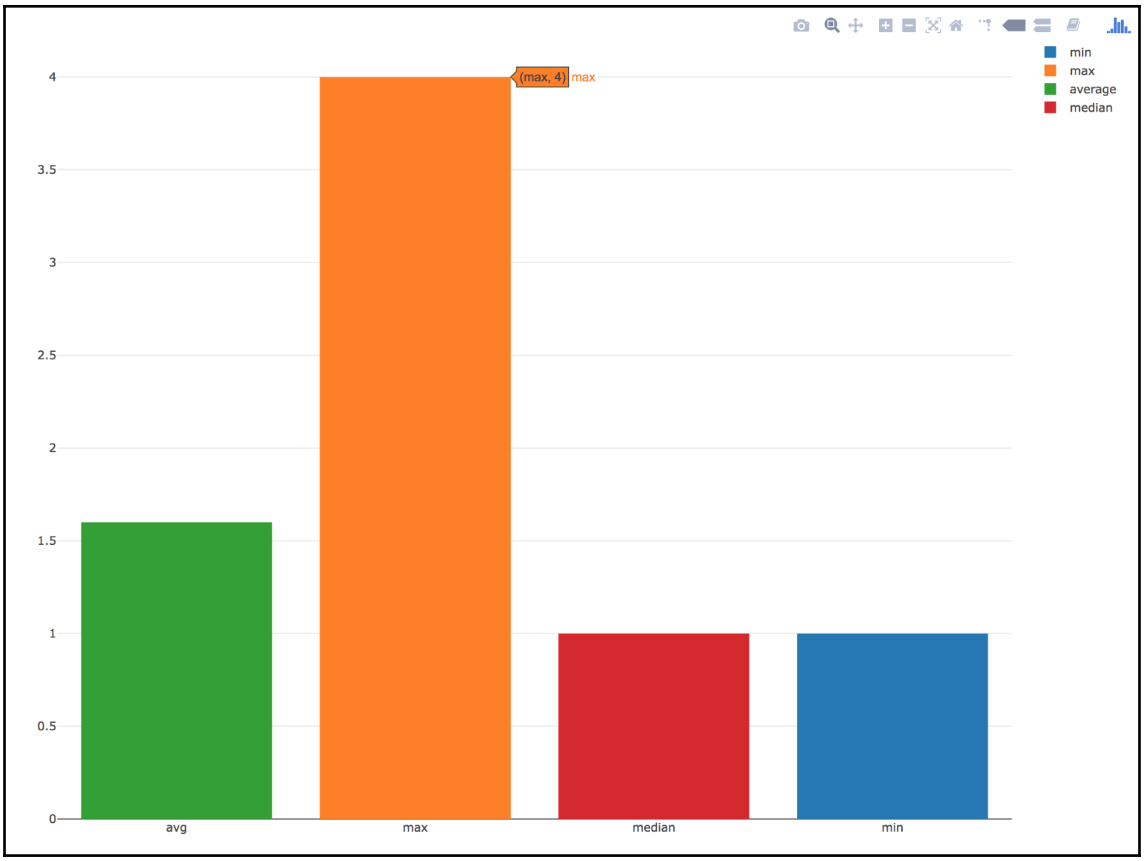

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/p[1]/span[1]/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]
0: text
assignedSlot: null
baseURI: "http://devveri.com/"
childNodes: NodeList []
data: "18 Ocak 2018"
firstChild: null
isConnected: true
lastChild: null
length: 12
nextElementSibling: null
nextSibling: null
nodeName: "#text"
nodeType: 3
nodeValue: "18 Ocak 2018"
ownerDocument: document
parentElement: span.date
parentNode: span.date
previousElementSibling: null
previousSibling: null
textContent: "18 Ocak 2018"
wholeText: "18 Ocak 2018"
__proto__: Text


Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/p[1]/span[4]/a/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]

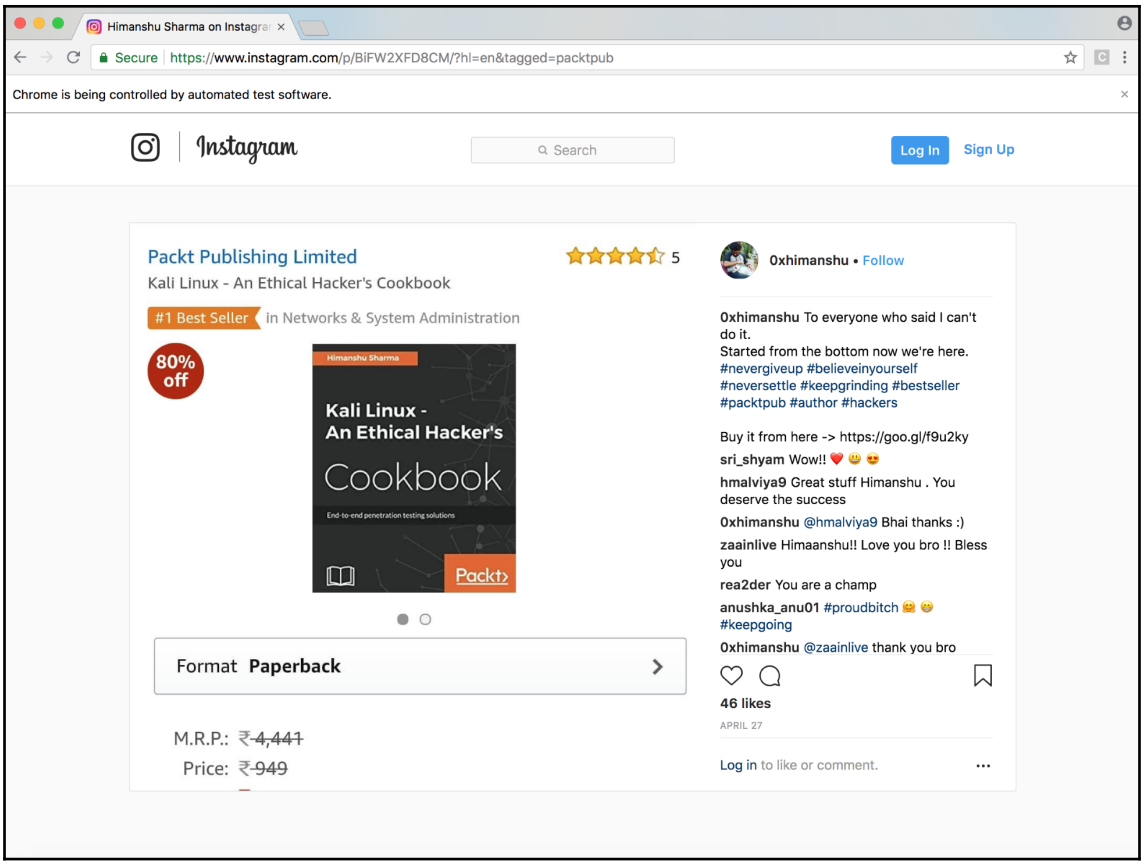

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[1]/div/p[1]/span[4]/a/text()')
(15) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]
0: text
assignedSlot: null
baseURI: "http://devveri.com/"
childNodes: NodeList []
data: "0"
firstChild: null
isConnected: true
lastChild: null
length: 1
nextElementSibling: null
nextSibling: null
nodeName: "#text"
nodeType: 3
nodeValue: "0"
ownerDocument: document
parentElement: a
parentNode: a
previousElementSibling: null
previousSibling: null
textContent: "0"
wholeText: "0"
__proto__: Text

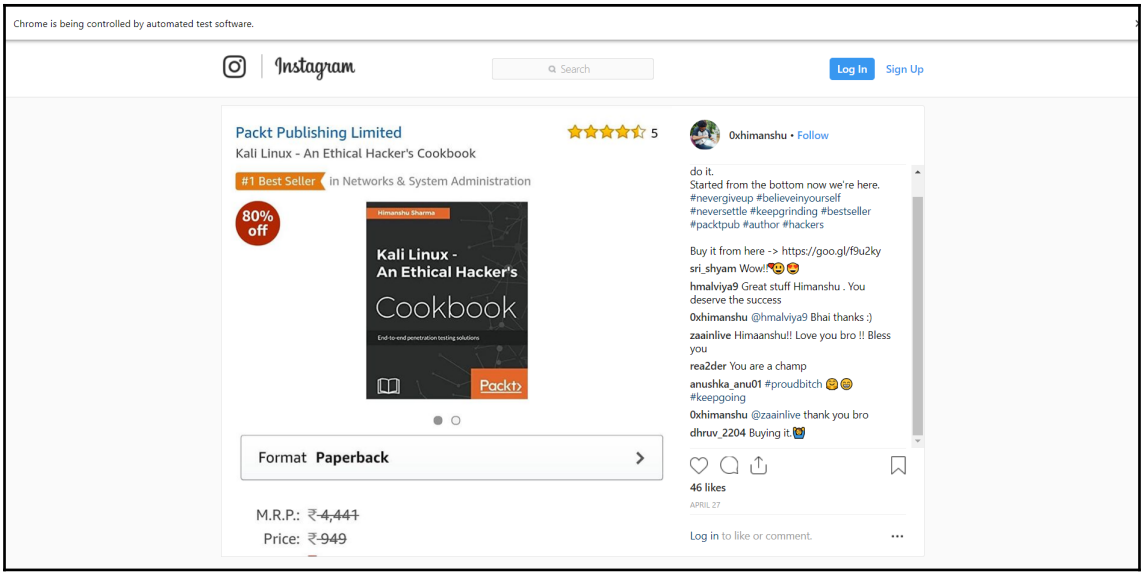

count
category
(ElasticSearch, 7)
0
1
2
3
4
5
6
7
8
Big Data
Cloud
docker
Doğal Dil İşleme
ElasticSearch
Graph
Haberler
Hadoop
HBase
Kitap
Lucene / Solr
Nosql
Ölçeklenebilirlik
Polyglot
Sunum
Veri Bilimi
Veri Madenciliği
Yapay

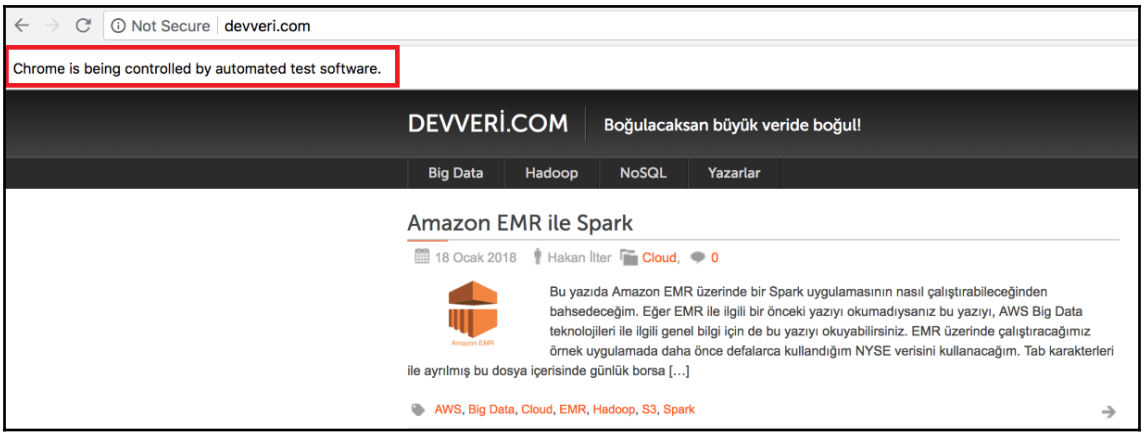

comment_count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
02 Kasım 2015
09 Aralık 2015
10 Ocak 2018
11 Şubat 2016
13 Ocak 2018
18 Şubat 2016
25 Kasım 2015
29 Mart 2016
date


min
max
average
median
4
3.5
3
2.5
2
1.5
1
0.5
0
(max, 4) max
avg
max
median
min

Chapter 4: Web Scraping with Rselenium


Himanshu Sharma on Instagra
Secure https://www.instagram.com/p/BiFW2XFD8CM/?hl=en&tagged=packtpub
Chrome is being controlled by automated test software.
Instagram
Search
Log In
Sign Up
Packt Publishing Limited
Kali Linux - An Ethical Hacker's Cookbook
#1 Best Seller in Networks & System Administration
80% off
Himanshu Sharma
Kali Linux - An Ethical Hacker's
Cookbook
End-to-end penetration testing solutions
Packt
Format Paperback
M.R.P.: ₹4,441
Price: ₹949
5
0xhimanshu • Follow
0xhimanshu To everyone who said I can't do it.
Started from the bottom now we're here.
#nevergiveup #believeinyourself #neversettle #keepgrinding #bestseller #packtpub #author #hackers
Buy it from here -> https://goo.gl/f9u2ky
sri_shyam Wow!!
hmalviya9 Great stuff Himanshu . You deserve the success
0xhimanshu @hmalviya9 Bhai thanks :)
zaainlive Himaanshu!! Love you bro !! Bless you
rea2der You are a champ
anushka_anu01 #proudbitch #keepgoing
0xhimanshu @zaainlive thank you bro
46 likes
APRIL 27
Log in to like or comment.


Chrome is being controlled by automated test software.
Instagram
Search
Log In
Sign Up
Packt Publishing Limited
Kali Linux - An Ethical Hacker's Cookbook
#1 Best Seller in Networks & System Administration
80% off
Himanshu Sharma
Kali Linux - An Ethical Hacker's Cookbook
End-to-end penetration testing solutions
Packt
Format Paperback
M.R.P.: ₹4,441
Price: ₹949
5
0xhimanshu • Follow
do it.
Started from the bottom now we're here.
#nevergiveup #believeinyourself
#neversettle #keepgrinding #bestseller
#packtpub #author #hackers
Buy it from here -> https://goo.gl/f9u2ky
sri_shyam Wow!
hmalviya9 Great stuff Himanshu . You deserve the success
0xhimanshu @hmalviya9 Bhai thanks :)
zaainlive Himaanshu!! Love you bro !! Bless you
rea2der You are a champ
anushka_anu01 #proudbitch
#keepgoing
0xhimanshu @zaainlive thank you bro
dhruv_2204 Buying it
46 likes
APRIL 27
Log in to like or comment.


Not Secure devveri.com
Chrome is being controlled by automated test software.
DEVVERİ.COM
Boğulacaksan büyük veride boğul!
Big Data
Hadoop
NoSQL
Yazarlar
Amazon EMR ile Spark
18 Ocak 2018
Hakan İlter
Cloud,
0
Amazon EMR
Bu yazıda Amazon EMR üzerinde bir Spark uygulamasının nasıl çalıştırabileceğinden bahsedeceğim. Eğer EMR ile ilgili bir önceki yazıyı okumadıysanız bu yazıyı, AWS Big Data teknolojileri ile ilgili genel bilgi için de bu yazıyı okuyabilirsiniz. EMR üzerinde çalıştıracağımız örnek uygulamada daha önce defalarca kullandığım NYSE verisini kullanacağım. Tab karakterleri ile ayrılmış bu dosya içerisinde günlük borsa […]
AWS, Big Data, Cloud, EMR, Hadoop, S3, Spark

Amazon EMR ile Spark

18 Ocak 2018 Hakan İlter Cloud, 0

Bu yazıda Amazon EMR üzerinde bir Spark uygulamasının nasıl çalıştırabileceğinden bahsedeceğim. Eğer EMR ile ilgili bir önceki yazıyı okumadıysanız bu yazıyı, AWS Big Data teknolojileri ile ilgili genel bilgi için de bu yazıyı okuyabilirsiniz.

EMR üzerinde çalıştıracağımız örnek uygulamada daha önce defalarca kullandığım NYSE verisini kullanacağım. Tab karakterleri ile ayrılmış bu dosya içerisinde günlük borsa işlemleri ile ilgili veriler bulunuyor. Spark 2.x ile birlikte gelen CSV desteği ile kolayca DataFrame haline getireceğimiz veriyi, S3 üzerinden okuyup üzerinde aggregation yapan bir SQL komutu çalıştırıp sonucu Parquet formatında tekrar S3 üzerine yazacağız.

Başlamadan önce işleyeceğimiz dosyayı indirip S3 üzerine atmamız gerekiyor. Bu işlemi elle AWS S3 arayüzünden yapabileceğiniz gibi komut satırından da yapabilirsiniz. Bunun için **awscli** uygulamasının yüklenmiş, ayarlarını yapmış olmanız ve S3 üzerindeki bucket'ı yaratmış olmanız gerekiyor.

```
1 wget https://s3.amazonaws.com/hw-sandbox/tutorial1/NYSE-2000-2001.tsv.gz
2 aws s3 cp NYSE-2000-2001.tsv.gz s3://datapyro-main/test/
```

EMR içerisinde kaynak yönetimi için **YARN** kullandığından Spark uygulamasında master parametresini "yarn" olarak belirtmek gerekiyor. Uygulamamız input parametresi ile S3 üzerinde okuyacağı veriyi, output parametresi ile de S3 üzerinde sonucun yazılacağı yeri alıyor.

AWS, Big Data, Cloud, EMR, Hadoop, S3, Spark

Amazon EMR

Amazon EMR

13 Ocak 2018 Hakan İlter Cloud, 0

Amazon EMR Nedir?

Amazon Elastic MapReduce (EMR), büyük veri işlemeyi kolaylaştırmak amacıyla Amazon tarafından yönetilen, içerisinde Hadoop, Spark gibi açık kaynaklı büyük veri teknolojilerini içeren bir servistir. Aslında temelde AWS üzerinde Hadoop kümesi kurmak için tek tek sunucuları açmak, gerekli yazılımları yüklemek gibi işlemleri otomatik olarak yapmaktadır. Bu sayede tek tık ile bir kümeyi kurabileceğiniz gibi, işiniz bittiğinde de yine tek tık ile kümeyi silebilirsiniz. Yani, bu esneklik sayesinde çok daha büyük bir kümeyi veriyi işleyeceğiniz zaman yaratıp, veriyi işledikten sonra da kümeyi kapatarak kaynakları boşuna kullanmadan hesaplı ve hızlı bir şekilde sonuca ulaşabilirsiniz.

Neden Amazon EMR Kullanırım?

Büyük veri teknolojilerine yeni başlayanların en çok sıkıntı çektiği konulardan birisi uygun çalışma ortamına sahip olmamak. Sanal makineler yardımıyla kendi bilgisayarımızda bu teknolojileri öğrenmek mümkün, ancak öğrendiklerinizi gerçek hayatta uygulamak için gerçek bir kümeye sahip olmak malesef çok mümkün olmuyor. Amazon EMR bu problemi ortadan kaldırıyor. İstediğiniz kümeyi kurup, kurcalayıp, kullandığınız kadar ödeyebilirsiniz

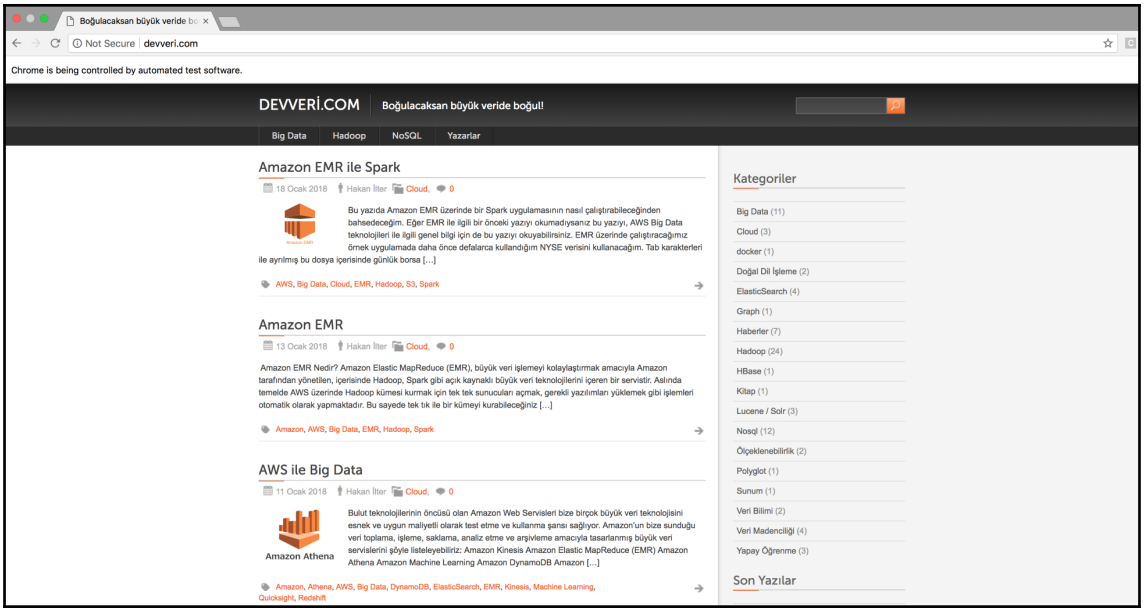
POC aşamasında sunucu satın alıp kaynak ayırmak ile uğraşmak yerine, EMR ile POC çalışmanızı ya da demonuzu yapabilirsiniz.

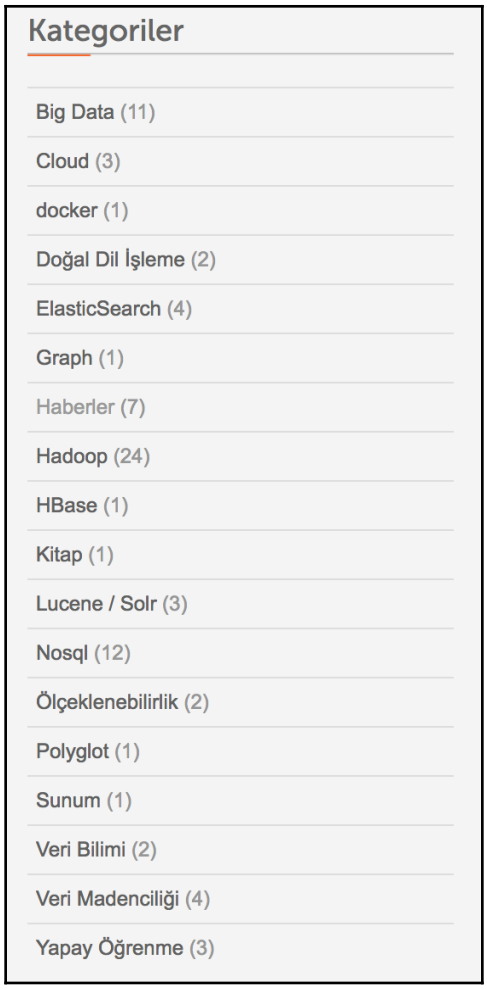
Firmalar için ise, kullandığınız kadar ödemek gerçekten anlamlı. Örneğin günde bir kere çalışan ve TB'larca veriyi işlemeyi gerektiren batch bir iş için sürekli ayakta duracak bir Hadoop kümesine ihtiyacınız yok. Ücretlendirme kümeyi oluşturan EC2 sunucularının büyüklük ve sayısına göre yapılıyor. Ayrıca spot instance kullanma şansınız da var.

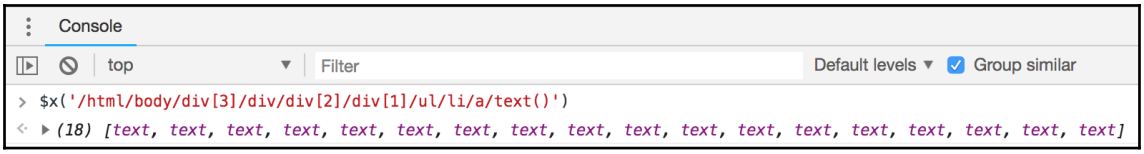
Spot instance kabaca, başkaları tarafından kiralanmış ancak boşta duran EC2 sunucularını normalden çok daha ucuza kullanmamızı sağlayan bir özellik. Bu avantaja rağmen ihtiyaç durumunda bu sunucuların sizden geri alınması dezavantajı var. Fakat büyük veri teknolojileri hata toleranslı olarak tasarlandıkları için, uygun tasarlanmış bir küme içerisindeki bazı sunucuların kapanması çalışan sistemi etkilemeyeceği için EMR ile spot instance kullanmak gerçekten iyi bir seçenek.

Kategoriler

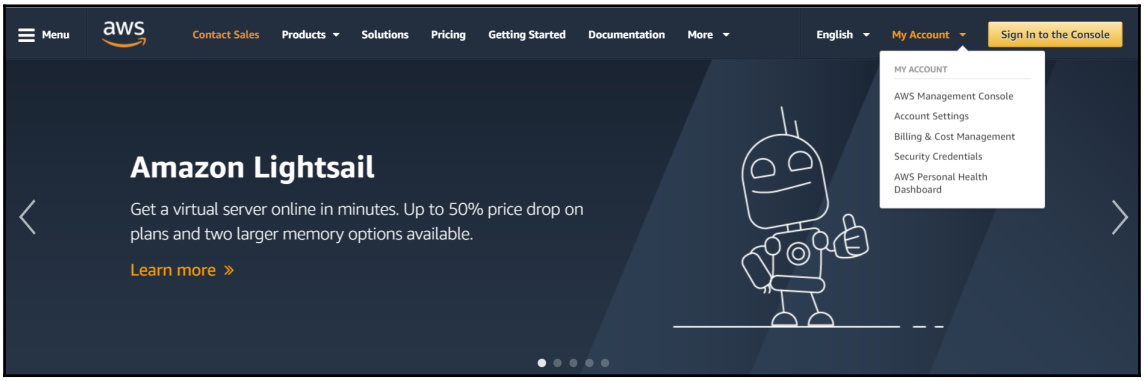
Big Data (11)
Cloud (3)
docker (1)
Doğal Dil İşleme (2)
ElasticSearch (4)
Graph (1)
Haberler (7)
Hadoop (24)
HBase (1)
Kitap (1)
Lucene / Solr (3)
Nosql (12)
Ölçeklenebilirlik (2)
Polyglot (1)
Sunum (1)
Veri Bilimi (2)
Veri Madenciliği (4)

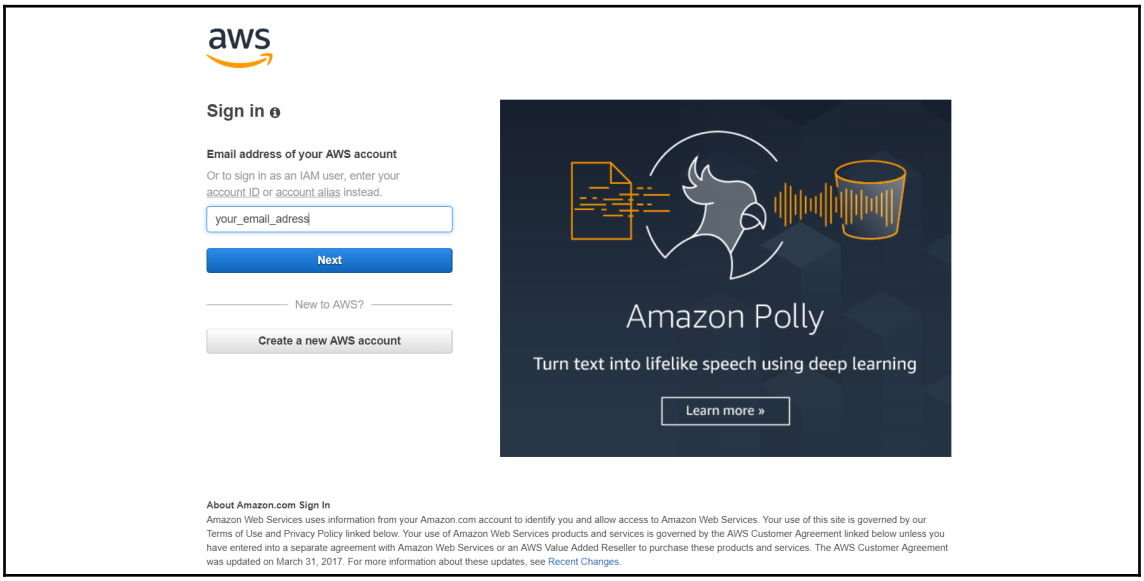

Boğulacaksan büyük veride bo
Not Secure devveri.com
Chrome is being controlled by automated test software.
DEVVERİ.COM
Boğulacaksan büyük veride boğul!
Big Data
Hadoop
NoSQL
Yazarlar
Amazon EMR ile Spark
18 Ocak 2018 Hakan İlter Cloud, 0
Bu yazıda Amazon EMR üzerinde bir Spark uygulamasının nasıl çalıştırabileceğinden bahsedeceğim. Eğer EMR ile ilgili bir önceki yazıyı okumadıysanız bu yazıyı, AWS Big Data teknolojileri ile ilgili genel bilgi için de bu yazıyı okuyabilirsiniz. EMR üzerinde çalıştıracağımız örnek uygulamada daha önce defalarca kullandığım NYSE verisini kullanacağım. Tab karakterleri ile ayrılmış bu dosya içerisinde günlük borsa […]
AWS, Big Data, Cloud, EMR, Hadoop, S3, Spark
Amazon EMR
13 Ocak 2018 Hakan İlter Cloud, 0
Amazon EMR Nedir? Amazon Elastic MapReduce (EMR), büyük veri işlemeyi kolaylaştırmak amacıyla Amazon tarafından yönetilen, içerisinde Hadoop, Spark gibi açık kaynaklı büyük veri teknolojilerini içeren bir servistir. Aslında temelde AWS üzerinde Hadoop kümesi kurmak için tek tek sunucuları açmak, gerekli yazılımları yüklemek gibi işlemleri otomatik olarak yapmaktadır. Bu sayede tek tık ile bir kümeyi kurabileceğiniz […]
Amazon, AWS, Big Data, EMR, Hadoop, Spark
AWS ile Big Data
11 Ocak 2018 Hakan İlter Cloud, 0
Amazon Athena
Bulut teknolojilerinin öncüsü olan Amazon Web Servisleri bize birçok büyük veri teknolojisini esnek ve uygun maliyetli olarak test etme ve kullanma şansı sağlıyor. Amazon'un bize sunduğu veri toplama, işleme, saklama, analiz etme ve arşivleme amacıyla tasarlanmış büyük veri servislerini şöyle listeleyebiliriz: Amazon Kinesis Amazon Elastic MapReduce (EMR) Amazon Athena Amazon Machine Learning Amazon DynamoDB Amazon […]
Amazon, Athena, AWS, Big Data, DynamoDB, ElasticSearch, EMR, Kinesis, Machine Learning, Quicksight, Redshift
Kategoriler
Big Data (11)
Cloud (3)
docker (1)
Doğal Dil İşleme (2)
ElasticSearch (4)
Graph (1)
Haberler (7)
Hadoop (24)
HBase (1)
Kitap (1)
Lucene / Solr (3)
Nosql (12)
Ölçeklenebilirlik (2)
Polyglot (1)
Sunum (1)
Veri Bilimi (2)
Veri Madenciliği (4)
Yapay Öğrenme (3)
Son Yazılar

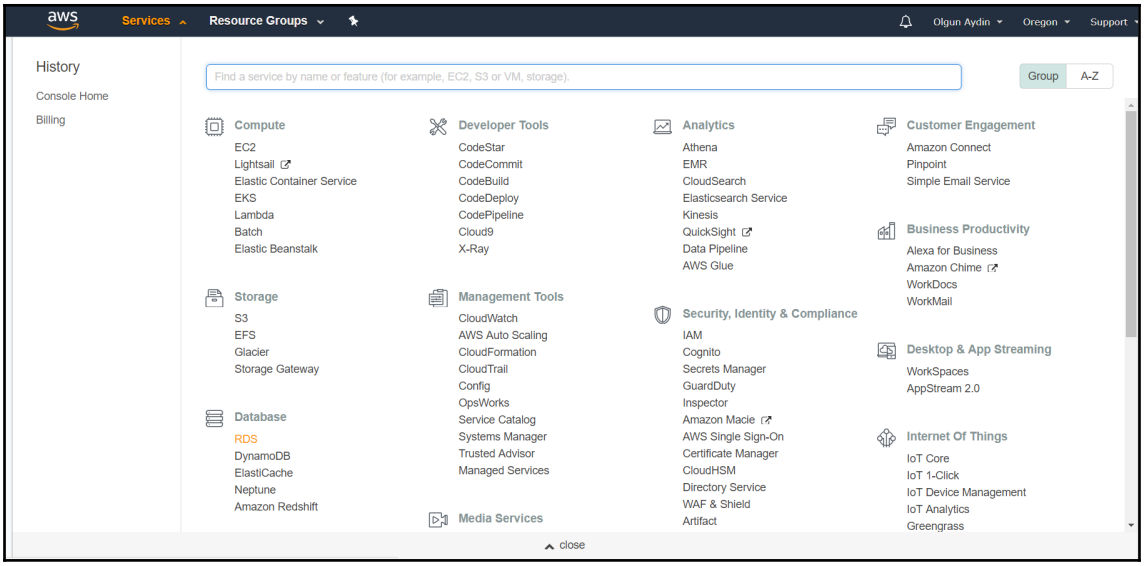

Kategoriler
Big Data (11)
Cloud (3)
docker (1)
Doğal Dil İşleme (2)
ElasticSearch (4)
Graph (1)
Haberler (7)
Hadoop (24)
HBase (1)
Kitap (1)
Lucene / Solr (3)
Nosql (12)
Ölçeklenebilirlik (2)
Polyglot (1)
Sunum (1)
Veri Bilimi (2)
Veri Madenciliği (4)
Yapay Öğrenme (3)

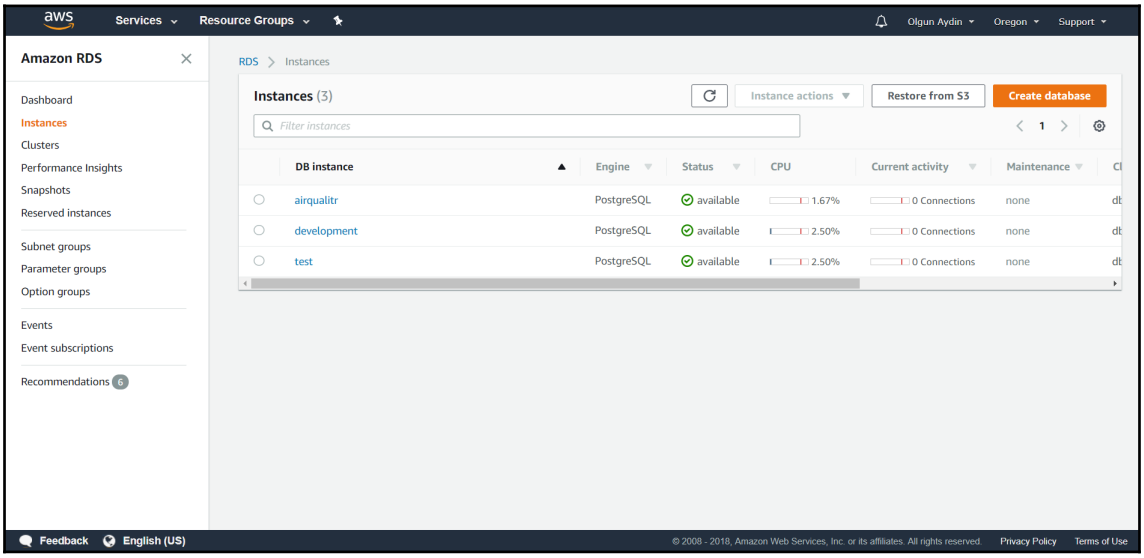

Console
top
Filter
Default levels
Group similar
$x('/html/body/div[3]/div/div[2]/div[1]/ul/li/a/text()')
(18) [text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text, text]

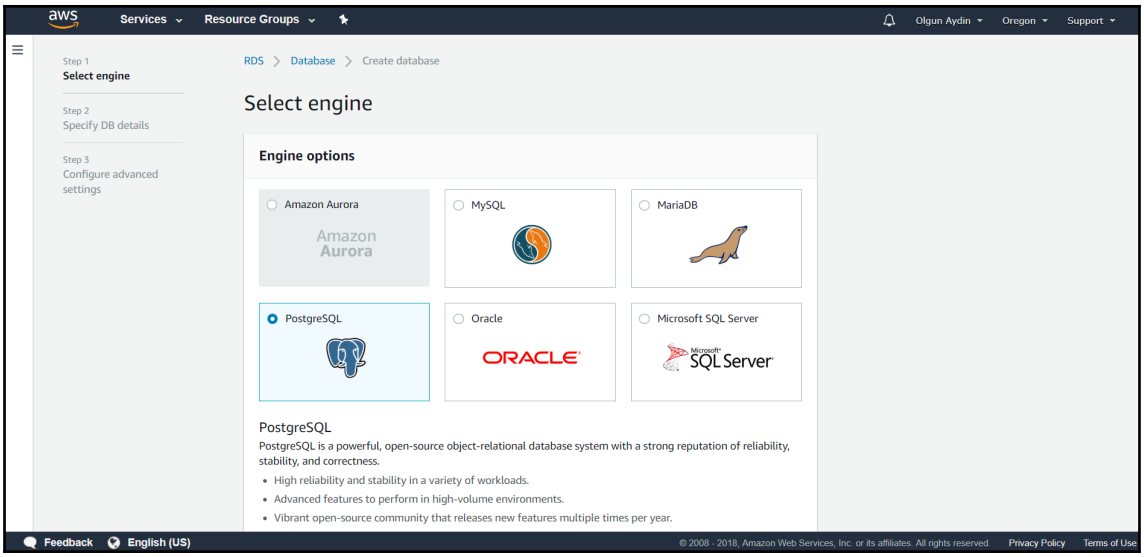
Chapter 5:
Storing Data and Creating Cronjob

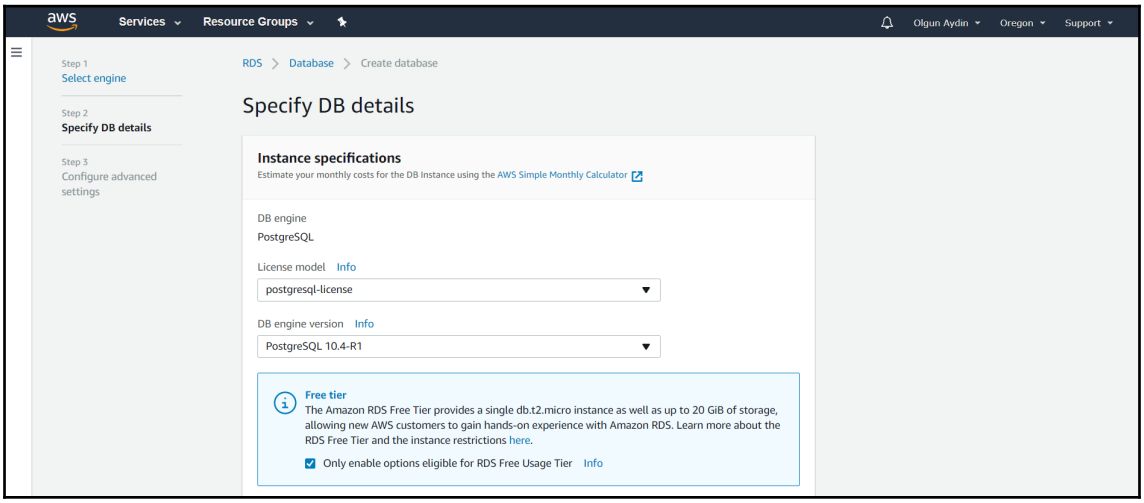


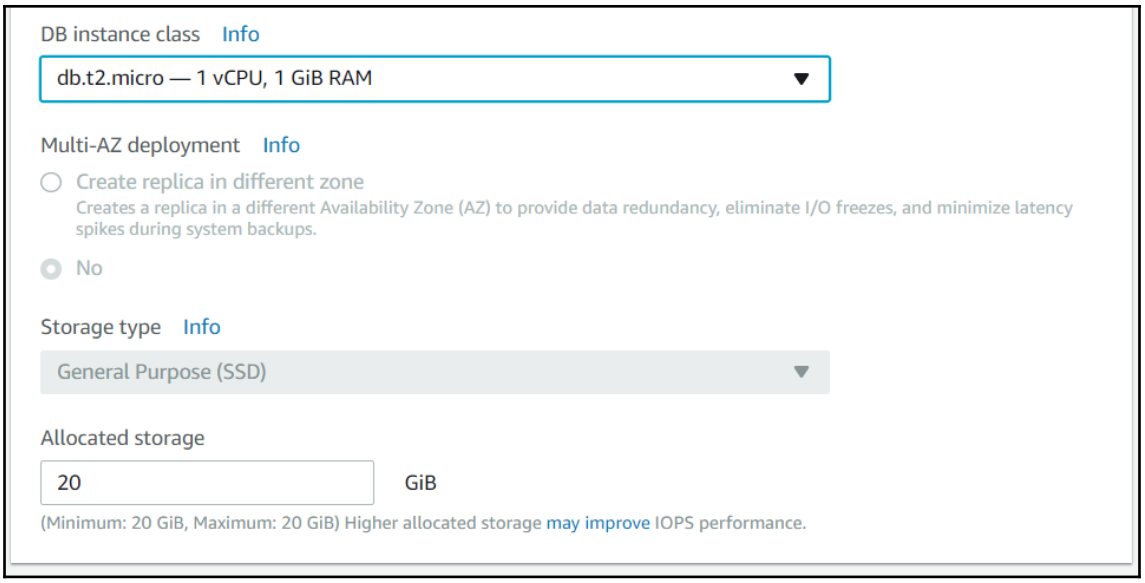


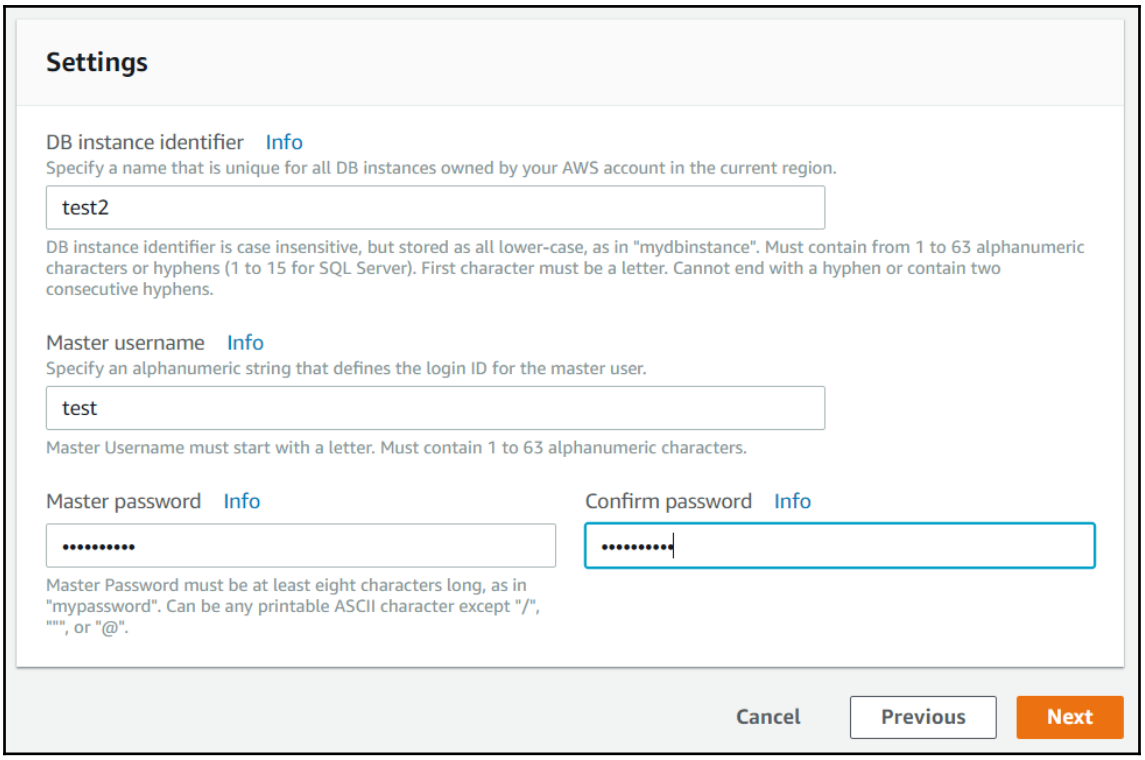

aws
Services
Resource Groups
Olgun Aydin
Oregon
Support
History
Console Home
Billing
Find a service by name or feature (for example, EC2, S3 or VM, storage).
Group
A-Z
Compute
EC2
Lightsail
Elastic Container Service
EKS
Lambda
Batch
Elastic Beanstalk
Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
Cloud9
X-Ray
Analytics
Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis
QuickSight
Data Pipeline
AWS Glue
Customer Engagement
Amazon Connect
Pinpoint
Simple Email Service
Business Productivity
Alexa for Business
Amazon Chime
WorkDocs
WorkMail
Storage
S3
EFS
Glacier
Storage Gateway
Management Tools
CloudWatch
AWS Auto Scaling
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Systems Manager
Trusted Advisor
Managed Services
Security, Identity & Compliance
IAM
Cognito
Secrets Manager
GuardDuty
Inspector
Amazon Macie
AWS Single Sign-On
Certificate Manager
CloudHSM
Directory Service
WAF & Shield
Artifact
Desktop & App Streaming
WorkSpaces
AppStream 2.0
Database
RDS
DynamoDB
ElastiCache
Neptune
Amazon Redshift
Media Services
Internet Of Things
IoT Core
IoT 1-Click
IoT Device Management
IoT Analytics
Greengrass
close


aws
Services
Resource Groups
Olgun Aydin
Oregon
Support
Amazon RDS
Dashboard
Instances
Clusters
Performance Insights
Snapshots
Reserved instances
Subnet groups
Parameter groups
Option groups
Events
Event subscriptions
Recommendations 6
RDS > Instances
Instances (3)
Instance actions
Restore from S3
Create database
Filter instances
1
DB instance
Engine
Status
CPU
Current activity
Maintenance
airqualitr
PostgreSQL
available
1.67%
0 Connections
none
development
PostgreSQL
available
2.50%
0 Connections
none
test
PostgreSQL
available
2.50%
0 Connections
none
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
Olgun Aydin
Oregon
Support
Step 1
Select engine
Step 2
Specify DB details
Step 3
Configure advanced settings
RDS > Database > Create database
Select engine
Engine options
Amazon Aurora
MySQL
MariaDB
PostgreSQL
Oracle
Microsoft SQL Server
PostgreSQL
PostgreSQL is a powerful, open-source object-relational database system with a strong reputation of reliability, stability, and correctness.
• High reliability and stability in a variety of workloads.
• Advanced features to perform in high-volume environments.
• Vibrant open-source community that releases new features multiple times per year.
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
Olgun Aydin
Oregon
Support
Step 1
Select engine
Step 2
Specify DB details
Step 3
Configure advanced settings
RDS > Database > Create database
Specify DB details
Instance specifications
Estimate your monthly costs for the DB Instance using the AWS Simple Monthly Calculator
DB engine
PostgreSQL
License model Info
postgresql-license
DB engine version Info
PostgreSQL 10.4-R1
Free tier
The Amazon RDS Free Tier provides a single db.t2.micro instance as well as up to 20 GiB of storage, allowing new AWS customers to gain hands-on experience with Amazon RDS. Learn more about the RDS Free Tier and the instance restrictions here.
Only enable options eligible for RDS Free Usage Tier Info


DB instance class Info
db.t2.micro — 1 vCPU, 1 GiB RAM
Multi-AZ deployment Info
Create replica in different zone
Creates a replica in a different Availability Zone (AZ) to provide data redundancy, eliminate I/O freezes, and minimize latency spikes during system backups.
No
Storage type Info
General Purpose (SSD)
Allocated storage
20
GiB
(Minimum: 20 GiB, Maximum: 20 GiB) Higher allocated storage may improve IOPS performance.


Settings
DB instance identifier Info
Specify a name that is unique for all DB instances owned by your AWS account in the current region.
test2
DB instance identifier is case insensitive, but stored as all lower-case, as in "mydbinstance". Must contain from 1 to 63 alphanumeric characters or hyphens (1 to 15 for SQL Server). First character must be a letter. Cannot end with a hyphen or contain two consecutive hyphens.
Master username Info
Specify an alphanumeric string that defines the login ID for the master user.
test
Master Username must start with a letter. Must contain 1 to 63 alphanumeric characters.
Master password Info
Confirm password Info
Master Password must be at least eight characters long, as in "mypassword". Can be any printable ASCII character except "/", """, or "@".
Cancel
Previous
Next

Database options

Database name Info

test2

If you do not specify a database name, Amazon RDS does not create a database.

Port Info

TCP/IP port the DB instance will use for application connections.

5432

DB parameter group Info

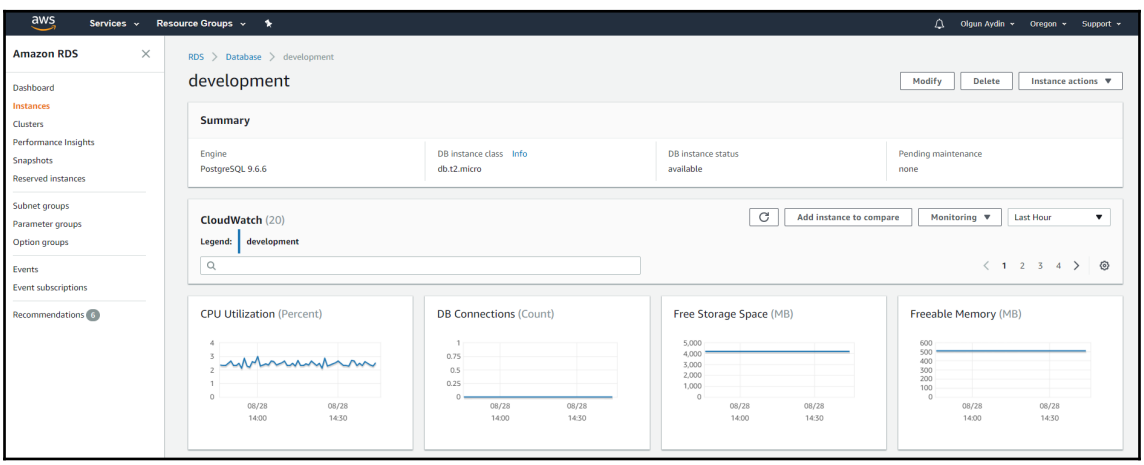
default.postgres10

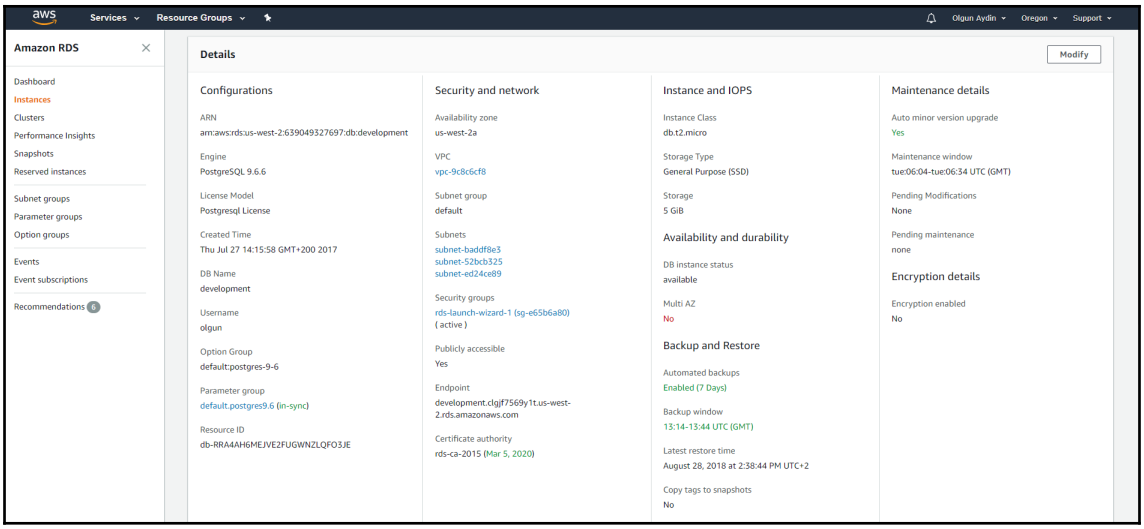
Option group Info

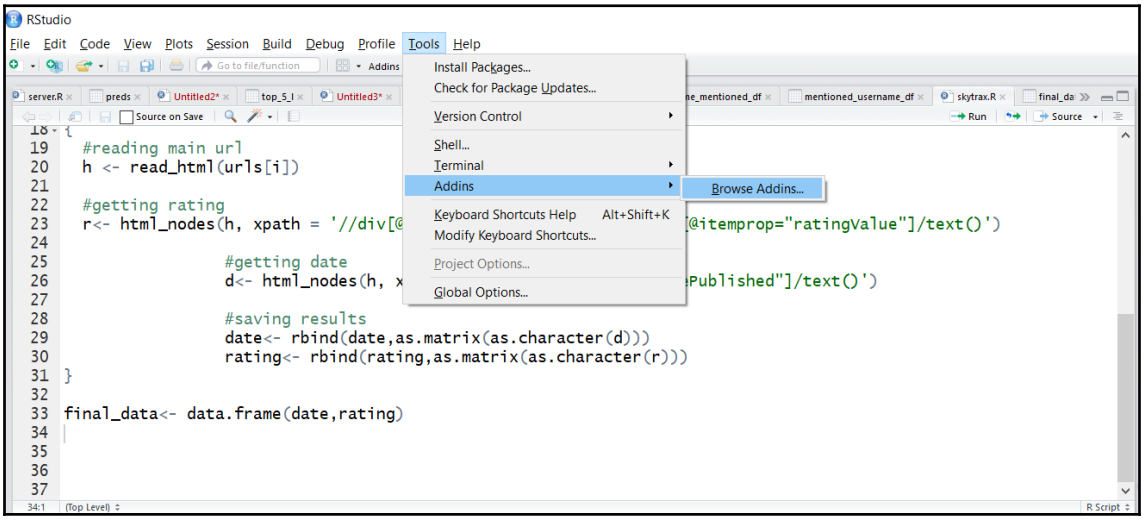
default:postgres-10

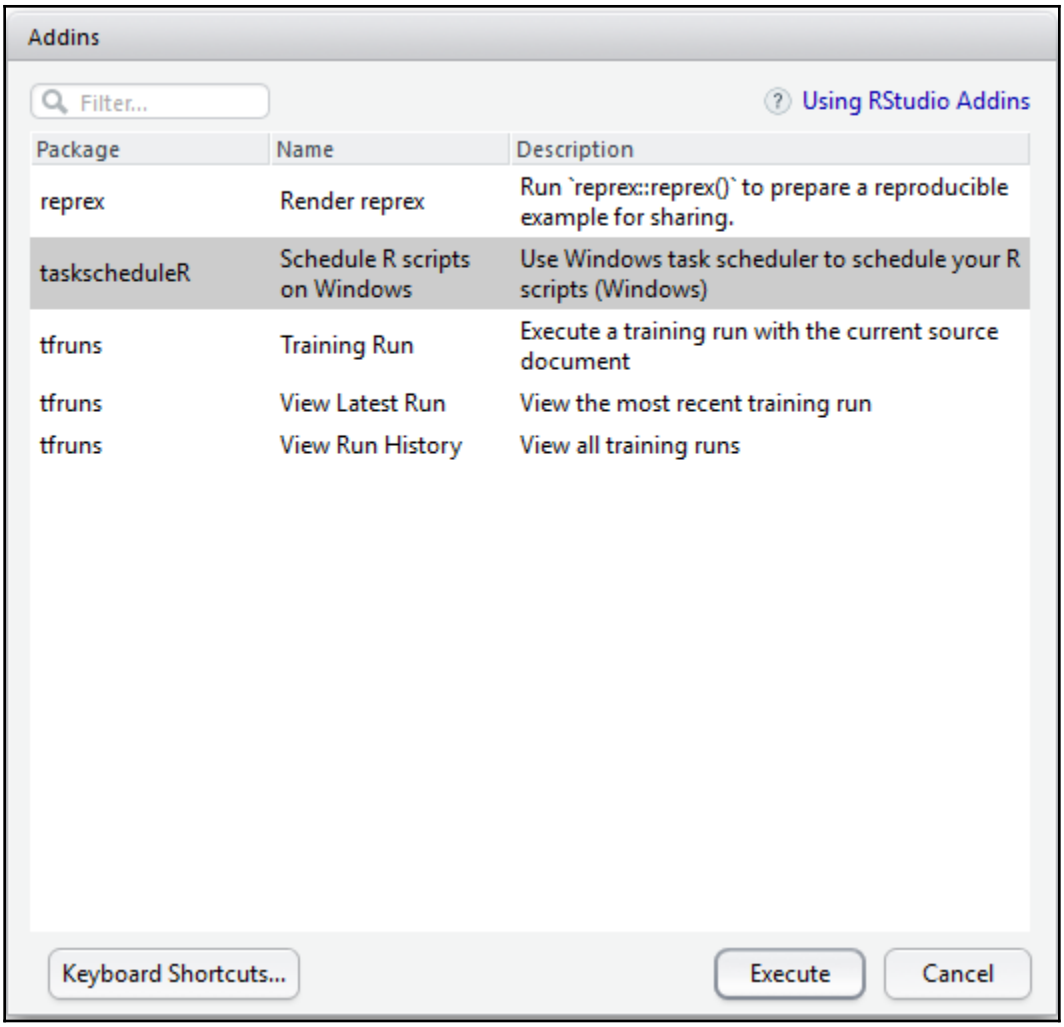
IAM DB authentication Info

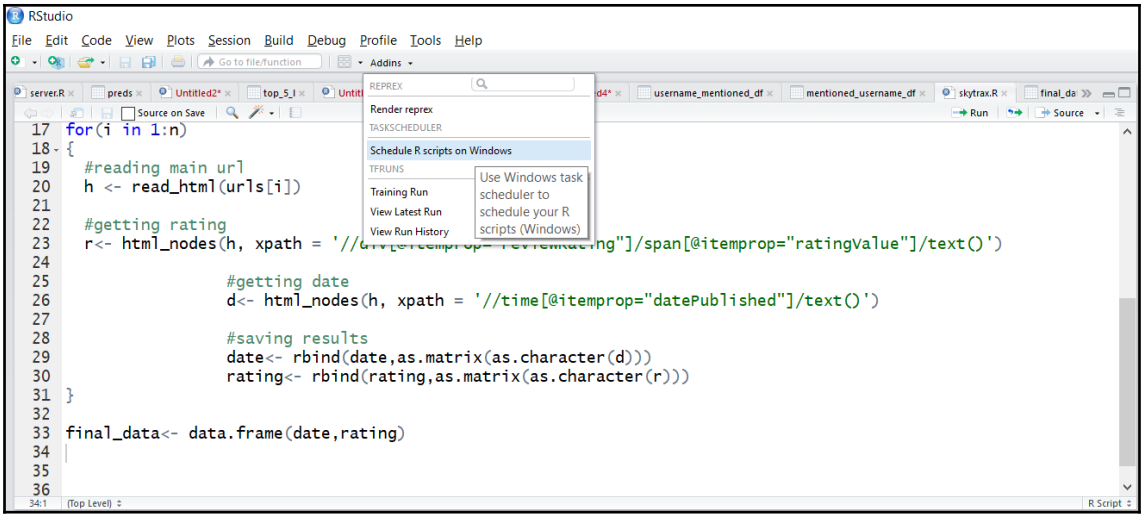
- Enable IAM DB authentication
  Manage your database user credentials through AWS IAM users and roles.
- Disable

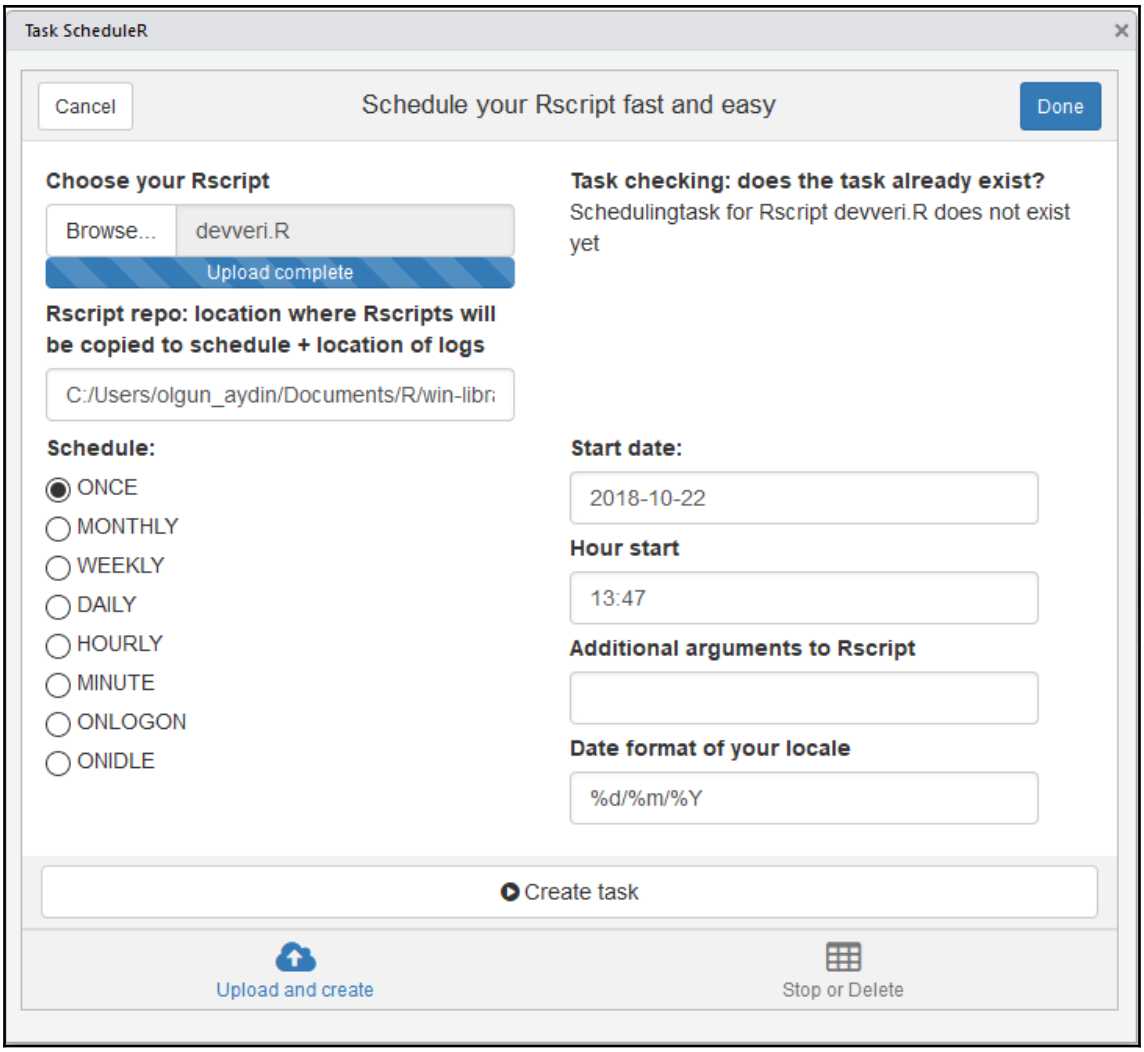



aws
Services
Resource Groups
Olgun Aydin
Oregon
Support
Amazon RDS
Dashboard
Instances
Clusters
Performance Insights
Snapshots
Reserved instances
Subnet groups
Parameter groups
Option groups
Events
Event subscriptions
Recommendations 6
Details
Modify
Configurations
ARN
arn:aws:rds:us-west-2:639049327697:db:development
Engine
PostgreSQL 9.6.6
License Model
Postgresql License
Created Time
Thu Jul 27 14:15:58 GMT+200 2017
DB Name
development
Username
olgun
Option Group
default:postgres-9-6
Parameter group
default.postgres9.6 (in-sync)
Resource ID
db-RRA4AH6MEJVE2FUGWNZLQFO3JE
Security and network
Availability zone
us-west-2a
VPC
vpc-9c8c6cf8
Subnet group
default
Subnets
subnet-baddf8e3
subnet-52bcb325
subnet-ed24ce89
Security groups
rds-launch-wizard-1 (sg-e65b6a80)
(active)
Publicly accessible
Yes
Endpoint
development.clqjf7569y1t.us-west-2.rds.amazonaws.com
Certificate authority
rds-ca-2015 (Mar 5, 2020)
Instance and IOPS
Instance Class
db.t2.micro
Storage Type
General Purpose (SSD)
Storage
5 GiB
Availability and durability
DB instance status
available
Multi AZ
No
Backup and Restore
Automated backups
Enabled (7 Days)
Backup window
13:14-13:44 UTC (GMT)
Latest restore time
August 28, 2018 at 2:38:44 PM UTC+2
Copy tags to snapshots
No
Maintenance details
Auto minor version upgrade
Yes
Maintenance window
tue:06:04-tue:06:34 UTC (GMT)
Pending Modifications
None
Pending maintenance
none
Encryption details
Encryption enabled
No


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Install Packages...
Check for Package Updates...
Version Control
Shell...
Terminal
Addins
Browse Addins...
Keyboard Shortcuts Help Alt+Shift+K
Modify Keyboard Shortcuts...
Project Options...
Global Options...
server.R
preds
Untitled2*
top_5_l
Untitled3*
mentioned_df
mentioned_username_df
skytrax.R
final_da
Source on Save
Run
Source
{
#reading main url
h <- read_html(urls[i])
#getting rating
r<- html_nodes(h, xpath = '//div[@
[@itemprop="ratingValue"]/text()')
#getting date
d<- html_nodes(h, x
Published"]/text()')
#saving results
date<- rbind(date,as.matrix(as.character(d)))
rating<- rbind(rating,as.matrix(as.character(r)))
}
final_data<- data.frame(date,rating)
34:1
(Top Level)
R Script


Addins
Filter...
Using RStudio Addins
Package
Name
Description
reprex
Render reprex
Run `reprex::reprex()` to prepare a reproducible example for sharing.
taskscheduleR
Schedule R scripts on Windows
Use Windows task scheduler to schedule your R scripts (Windows)
tfruns
Training Run
Execute a training run with the current source document
tfruns
View Latest Run
View the most recent training run
tfruns
View Run History
View all training runs
Keyboard Shortcuts...
Execute
Cancel


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
server.R
preds
Untitled2*
top_5_l
username_mentioned_df
mentioned_username_df
skytrax.R
Source on Save
Run
Source
REPREX
Render reprex
TASKSCHEDULER
Schedule R scripts on Windows
TFRUNS
Training Run
View Latest Run
View Run History
Use Windows task scheduler to schedule your R scripts (Windows)
for(i in 1:n)
{
#reading main url
h <- read_html(urls[i])
#getting rating
r<- html_nodes(h, xpath = '//
ng"]/span[@itemprop="ratingValue"]/text()')
#getting date
d<- html_nodes(h, xpath = '//time[@itemprop="datePublished"]/text()')
#saving results
date<- rbind(date,as.matrix(as.character(d)))
rating<- rbind(rating,as.matrix(as.character(r)))
}
final_data<- data.frame(date,rating)
34:1
(Top Level)
R Script


Task ScheduleR
Cancel
Schedule your Rscript fast and easy
Done
Choose your Rscript
Browse...
devveri.R
Upload complete
Rscript repo: location where Rscripts will be copied to schedule + location of logs
C:/Users/olgun_aydin/Documents/R/win-libr
Schedule:
ONCE
MONTHLY
WEEKLY
DAILY
HOURLY
MINUTE
ONLOGON
ONIDLE
Task checking: does the task already exist?
Schedulingtask for Rscript devveri.R does not exist yet
Start date:
2018-10-22
Hour start
13:47
Additional arguments to Rscript
Date format of your locale
%d/%m/%Y
Create task
Upload and create
Stop or Delete