

Data Engineering 4 - Home Assignment

Patrik Szigeti (CEU ID: 121536)

04/12/2020

Tech Setup (Instance ID: i-0026e1085400d6372)

I set up my EC2 instance with the following configurations:

- de4-week3 AMI:

The screenshot shows the AWS Management Console interface for the 'Launch instance wizard'. The browser address bar indicates the URL: `eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#LaunchInstanceWizard:`. The console header shows the user is logged in as 'szigeti_patrik @ ceu' in the 'Ireland' region. The navigation bar includes 'Services', 'Resource Groups', and 'EC2'. The wizard progress bar shows steps: 1. Choose AMI, 2. Choose Instance Type, 3. Configure Instance, 4. Add Storage, 5. Add Tags, 6. Configure Security Group, 7. Review. The current step is 'Step 1: Choose an Amazon Machine Image (AMI)'. Below the step title, a description states: 'An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.' A search bar is present with the placeholder text 'Search for an AMI by entering a search term e.g. "Windows"'. On the left, a sidebar shows 'Quick Start' and 'My AMIs' with filters for 'Ownership' (Owned by me, Shared with me) and 'Architecture' (32-bit (x86), 64-bit (x86), 64-bit (Arm)). The main content area displays a list of AMIs. The 'de4-week3' AMI is highlighted with a dashed border. The AMI list includes: 'stock-ubuntu-18_04-but-ssh-on-port-8000 - ami-05e29314ed71317fa', 'de4-week3 - ami-071c5ca9e5c81aed1', 'Ubuntu - SSH on port 8787 - ami-083d3d7f15e5cdcd3', and 'de4-week2 - ami-0ab0a49578237fe0c'. Each AMI entry shows its name, description, root device type, virtualization type, owner, and ENA status, along with a 'Select' button. The bottom of the screenshot shows the Windows taskbar with various application icons and the system tray displaying the date and time as '11:53 2020. 04. 12.'.

- t3.small instance type:

Launch instance wizard | EC2 M... x

eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#LaunchInstanceWizard:

Services Resource Groups EC2

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 2: Choose an Instance Type

Instance Type	Architecture	VCPU	Memory (GiB)	Storage	Network	Price (USD)	On-Demand Price (USD)
General purpose	t3a.xlarge	4	16	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3a.2xlarge	8	32	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.nano	2	0.5	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.micro	2	1	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.small	2	2	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.medium	2	4	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.large	2	8	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.xlarge	4	16	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	t3.2xlarge	8	32	EBS only	Yes	Up to 5 Gigabit	Yes
General purpose	m5ad.large	2	8	1 x 75 (SSD)	Yes	Up to 10 Gigabit	Yes
General purpose	m5ad.xlarge	4	16	1 x 150 (SSD)	Yes	Up to 10 Gigabit	Yes
General purpose	m5ad.2xlarge	8	32	1 x 300 (SSD)	Yes	Up to 10 Gigabit	Yes

Cancel Previous Review and Launch Next: Configure Instance Details

Feedback English (US) © 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use 11:54 2020.04.12

- gergely-week2 IAM role:

Launch instance wizard | EC2 M... x

eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#LaunchInstanceWizard:

Services Resource Groups EC2

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances 1 Launch into Auto Scaling Group

Purchasing option Request Spot instances

Network vpc-cf69a3a9 (default) Create new VPC

Subnet No preference (default subnet in any Availability Zone) Create new subnet

Auto-assign Public IP Use subnet setting (Enable)

Placement group Add instance to placement group

Capacity Reservation Open Create new Capacity Reservation

IAM role gergely-week2 Create new IAM role

CPU options Specify CPU options

Shutdown behavior Stop

Enable termination protection Protect against accidental termination

Cancel Previous Review and Launch Next: Add Storage

Feedback English (US) © 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use 11:54 2020.04.12

- Added ports 8000 for alternate SSH, 8787 for RStudio and 8080 for Jenkins:

Launch instance wizard | EC2 M5 x

eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#LaunchInstanceWizard:

Services Resource Groups EC2

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. [Learn more](#) about Amazon EC2 security groups.

Assign a security group: ☒ Create a new security group ☐ Select an existing security group

Security group name: patrik-szigeti-assignment

Description: patrik-szigeti-assignment created 2020-04-12T11:52:58.483+02:00

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	Custom 0.0.0.0/0	ssh
Custom TCP F	TCP	8000	Custom 0.0.0.0/0	alternate ssh
Custom TCP F	TCP	8787	Custom 0.0.0.0/0	rstudio
Custom TCP F	TCP	8080	Custom 0.0.0.0/0	jenkins

Add Rule

Warning
Rules with source of 0.0.0.0/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

Cancel Previous **Review and Launch**

Feedback English (US) © 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use 11:55 2020.04.12

- patrik-szigeti-assignment security group:

Launch instance wizard | EC2 M5 x

eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#LaunchInstanceWizard:

Services Resource Groups EC2

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 7: Review Instance Launch

de4-week3 - ami-071c5ca9e5c81aed1
Root Device Type: ebs Virtualization type: hvm

Instance Type [Edit instance type](#)

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t3.small	Variable	2	2	EBS only	Yes	Up to 5 Gigabit

Security Groups [Edit security groups](#)

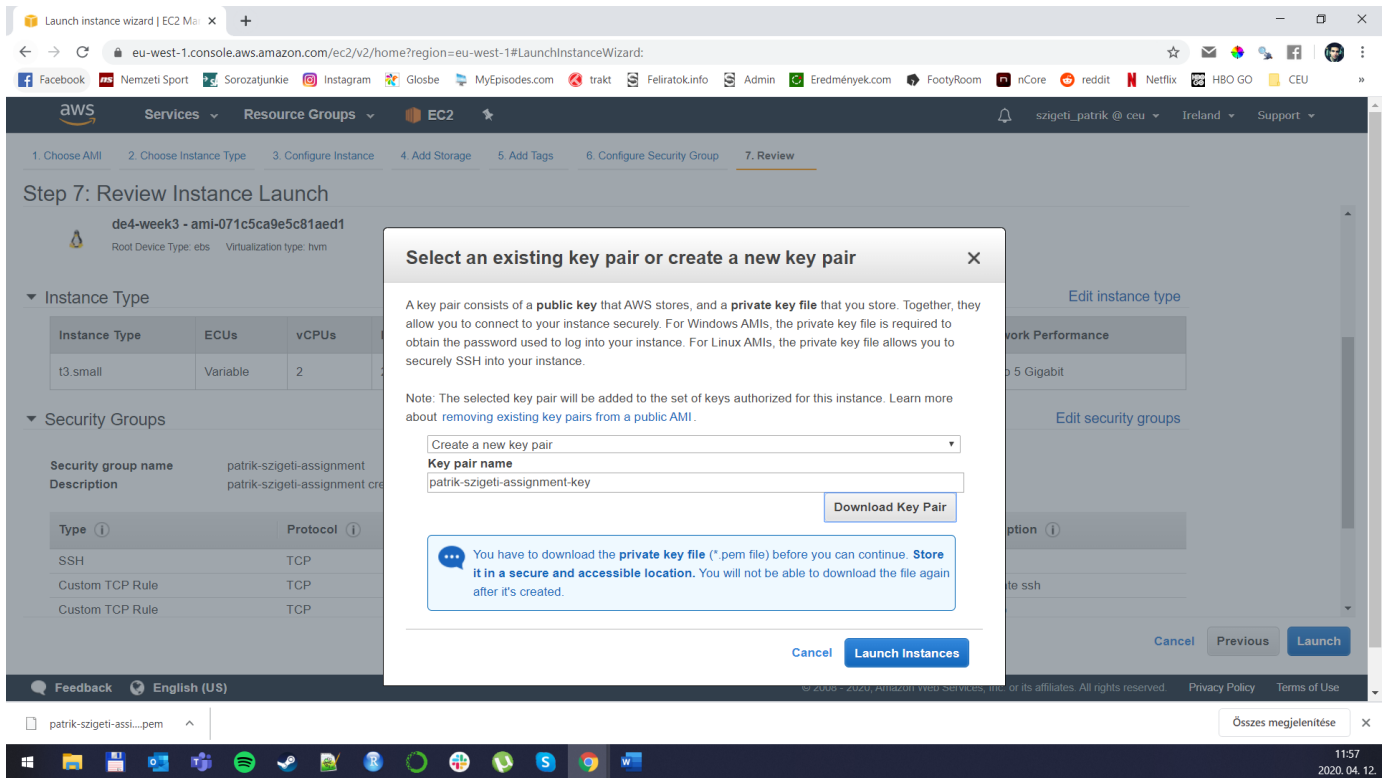
Security group name: patrik-szigeti-assignment
Description: patrik-szigeti-assignment created 2020-04-12T11:52:58.483+02:00

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	0.0.0.0/0	ssh
Custom TCP Rule	TCP	8000	0.0.0.0/0	alternate ssh
Custom TCP Rule	TCP	8787	0.0.0.0/0	rstudio
Custom TCP Rule	TCP	8080	0.0.0.0/0	jenkins

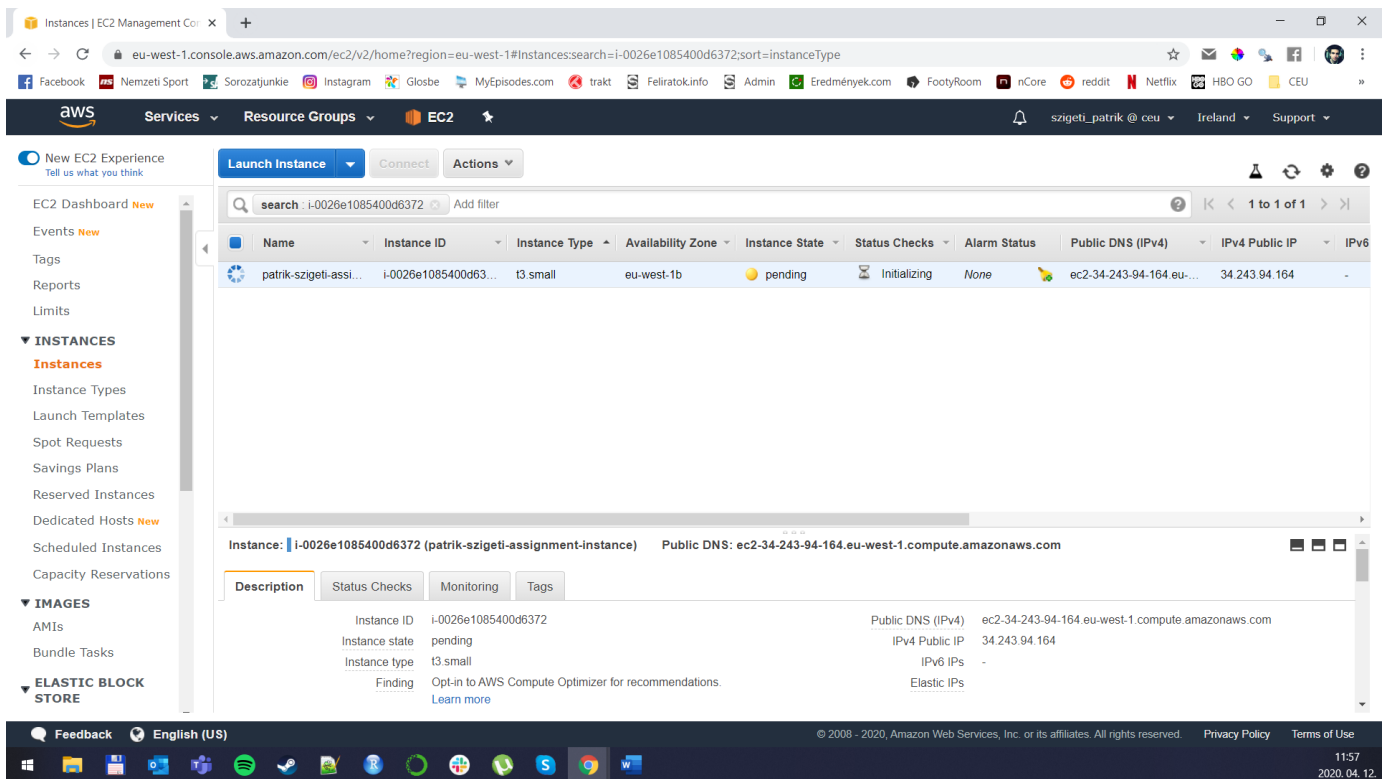
Cancel Previous **Launch**

Feedback English (US) © 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use 11:56 2020.04.12

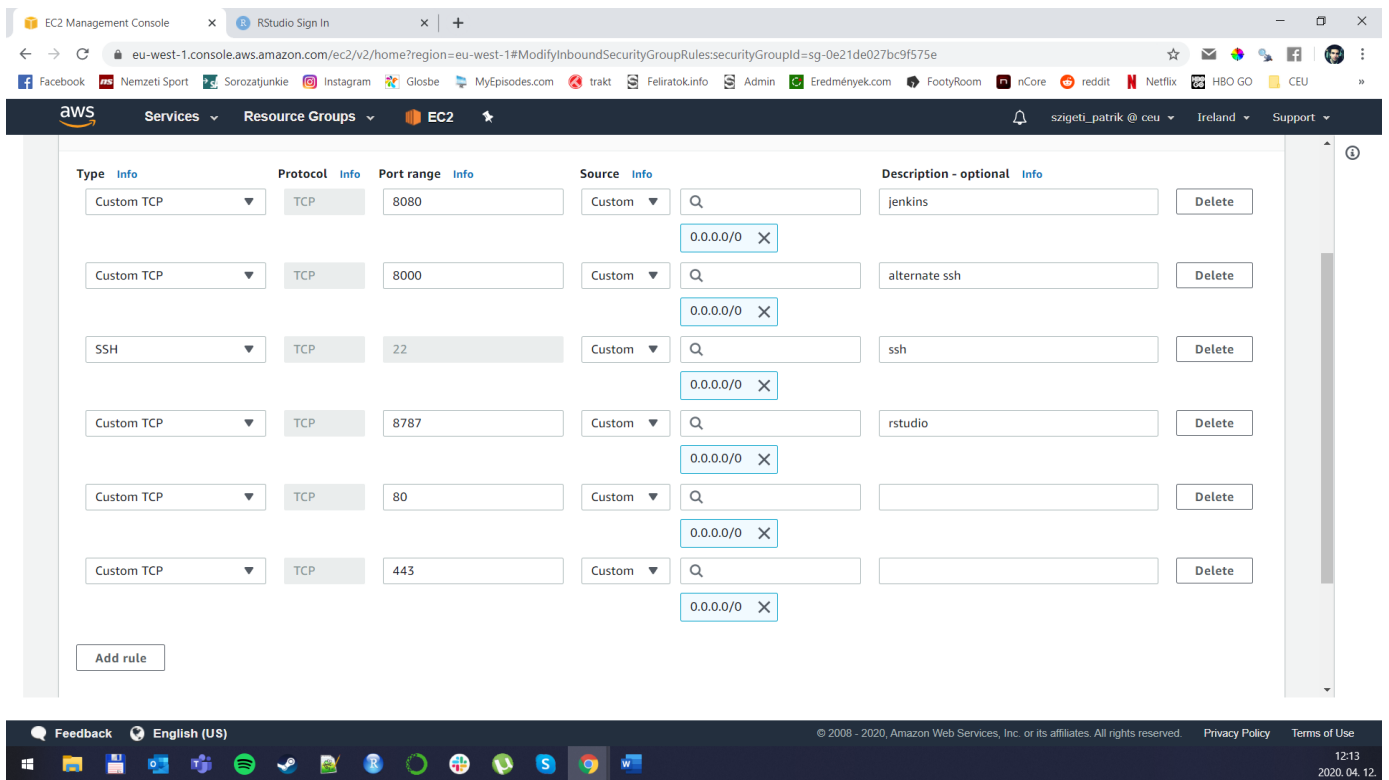
- patrik-szigeti-assignment-key newly created key pair:



- I named my instance `patrik-szigeti-assignment-instance`, and its Instance ID is `i-0026e1085400d6372`:

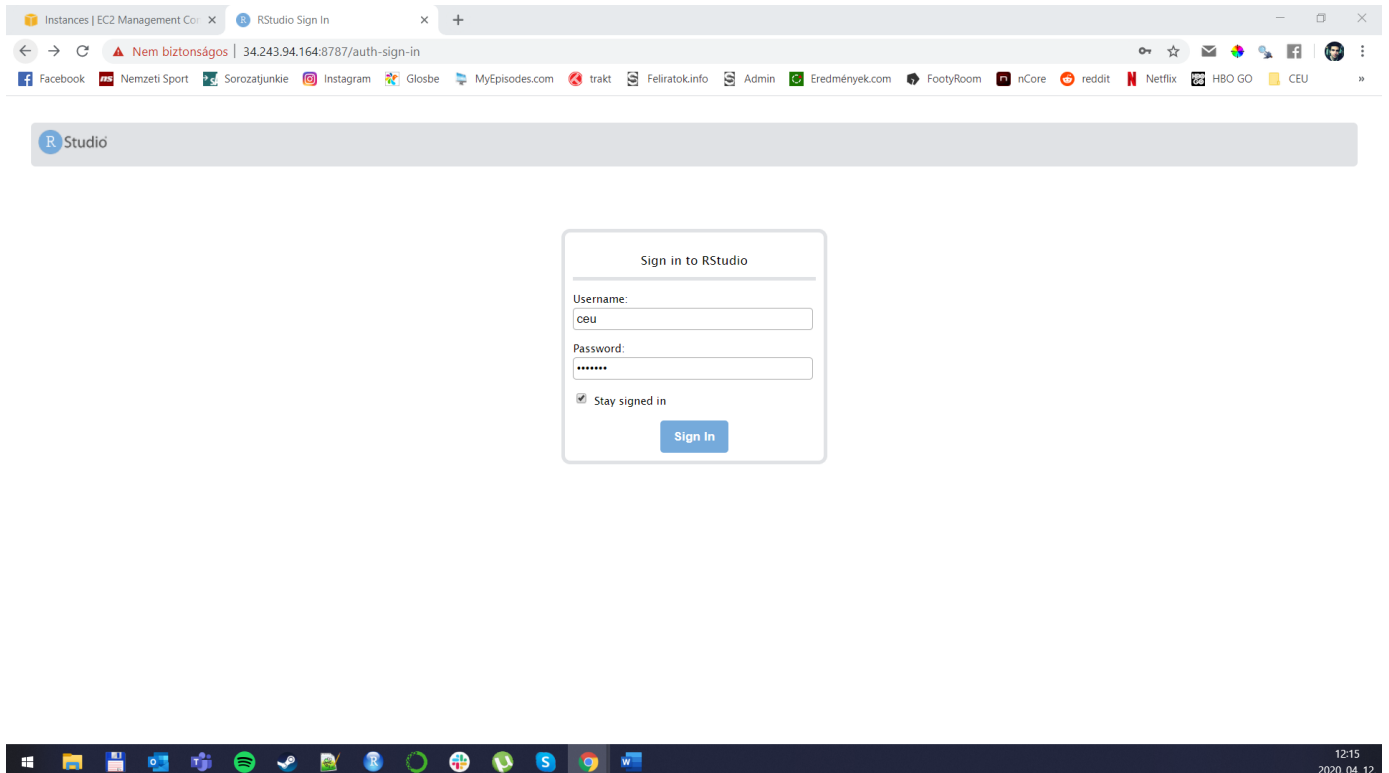


- I also included ports 80 and 443 so that the shortcuts would work:

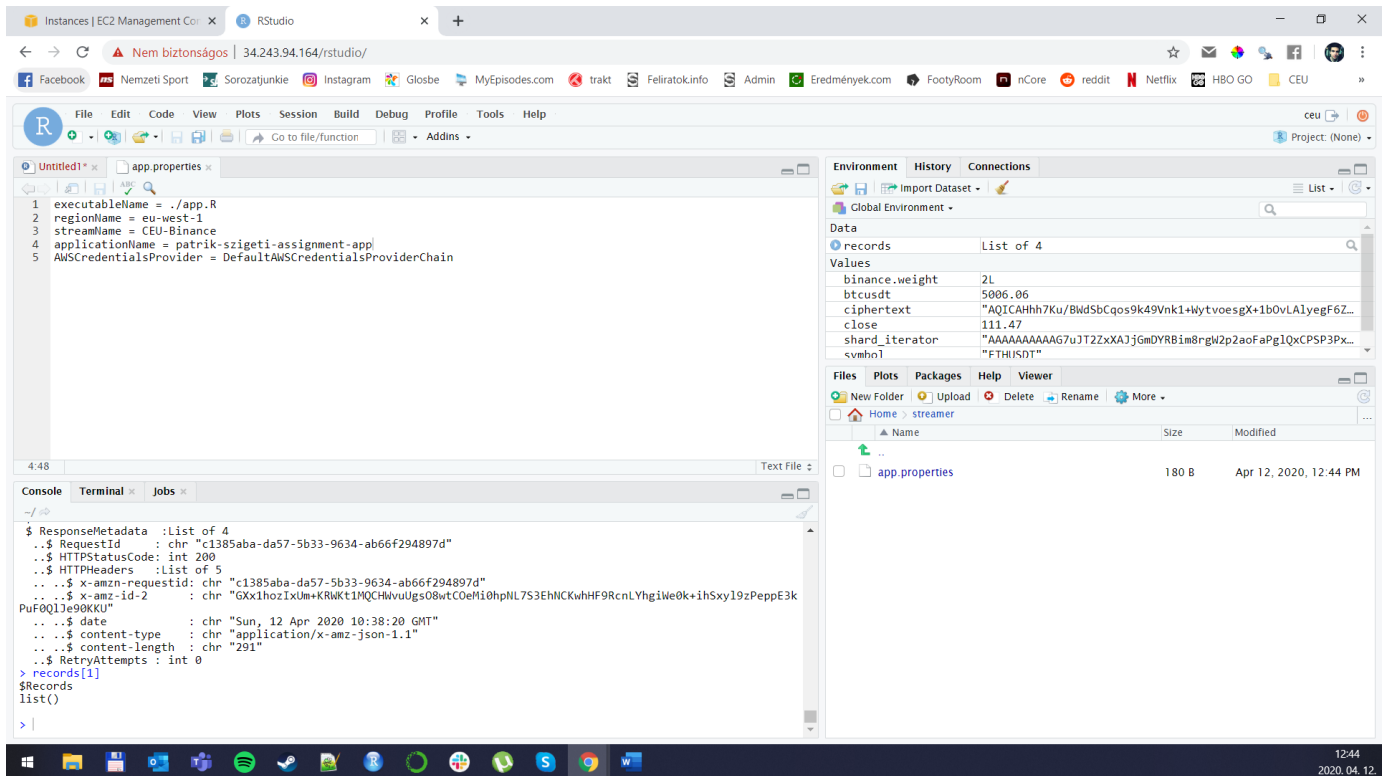


Stream Processing Application

The instance's setup was successful, logging in to RStudio with username `ceu` and password `ceudata` :

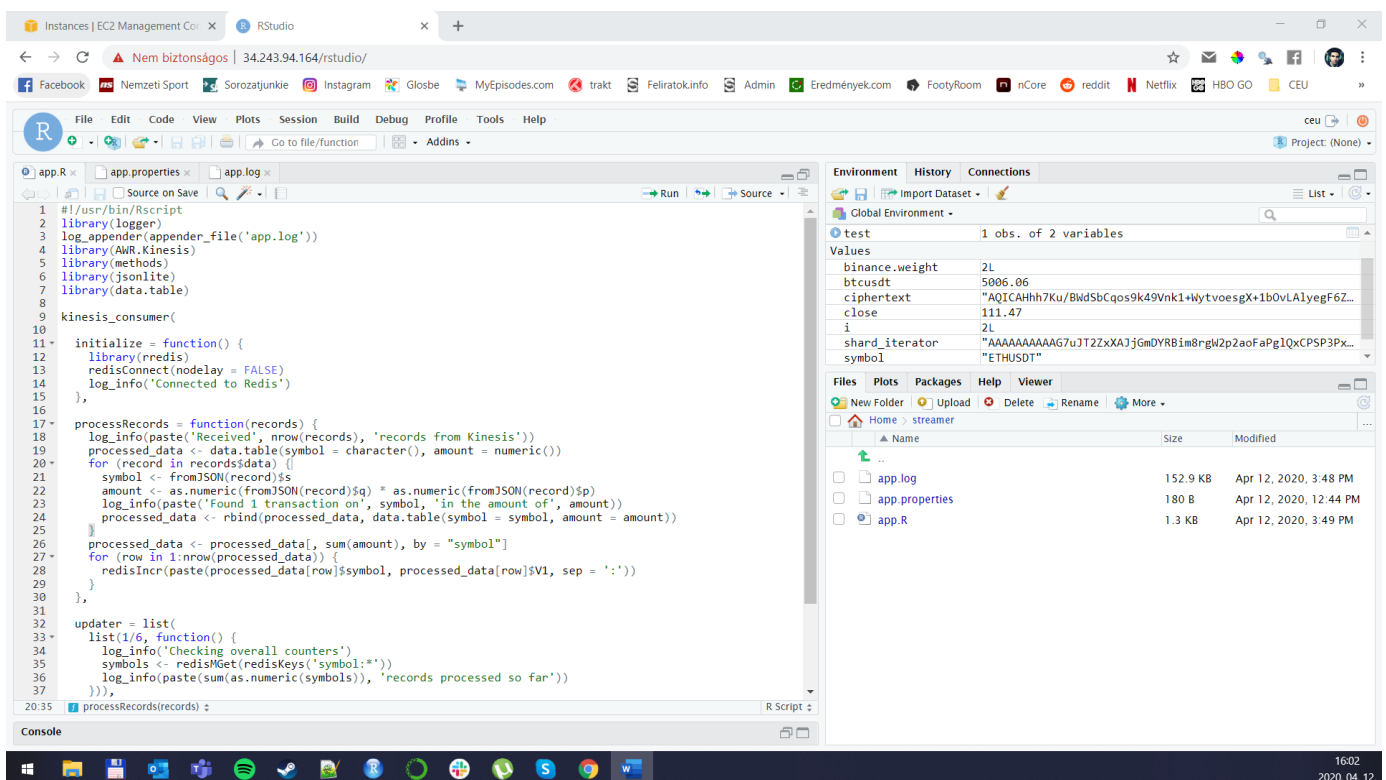


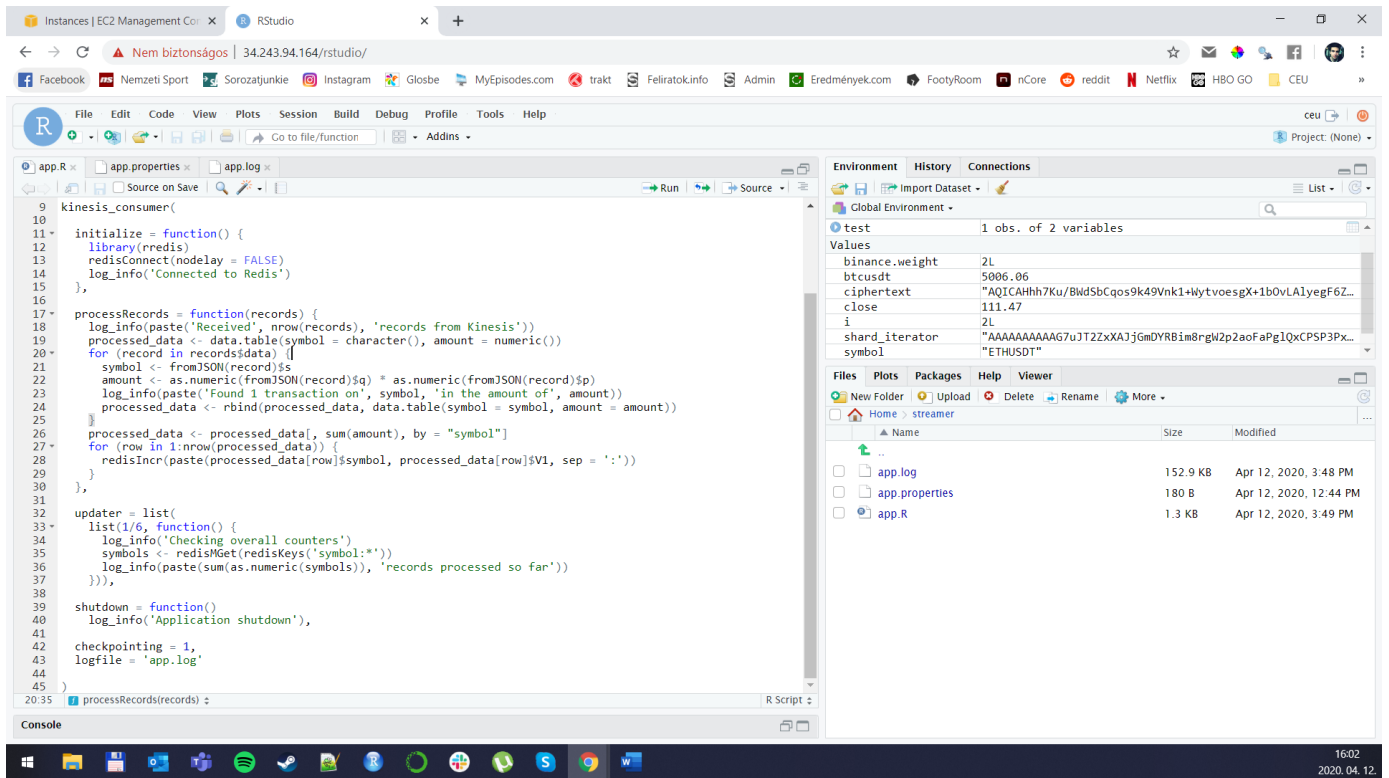
I set up an `app.properties` file in the streamer folder, set the stream name to `CEU-Binance` , the region to Ireland (`eu-west-1`) and named my application `patrik-szigeti-assignment-app` :



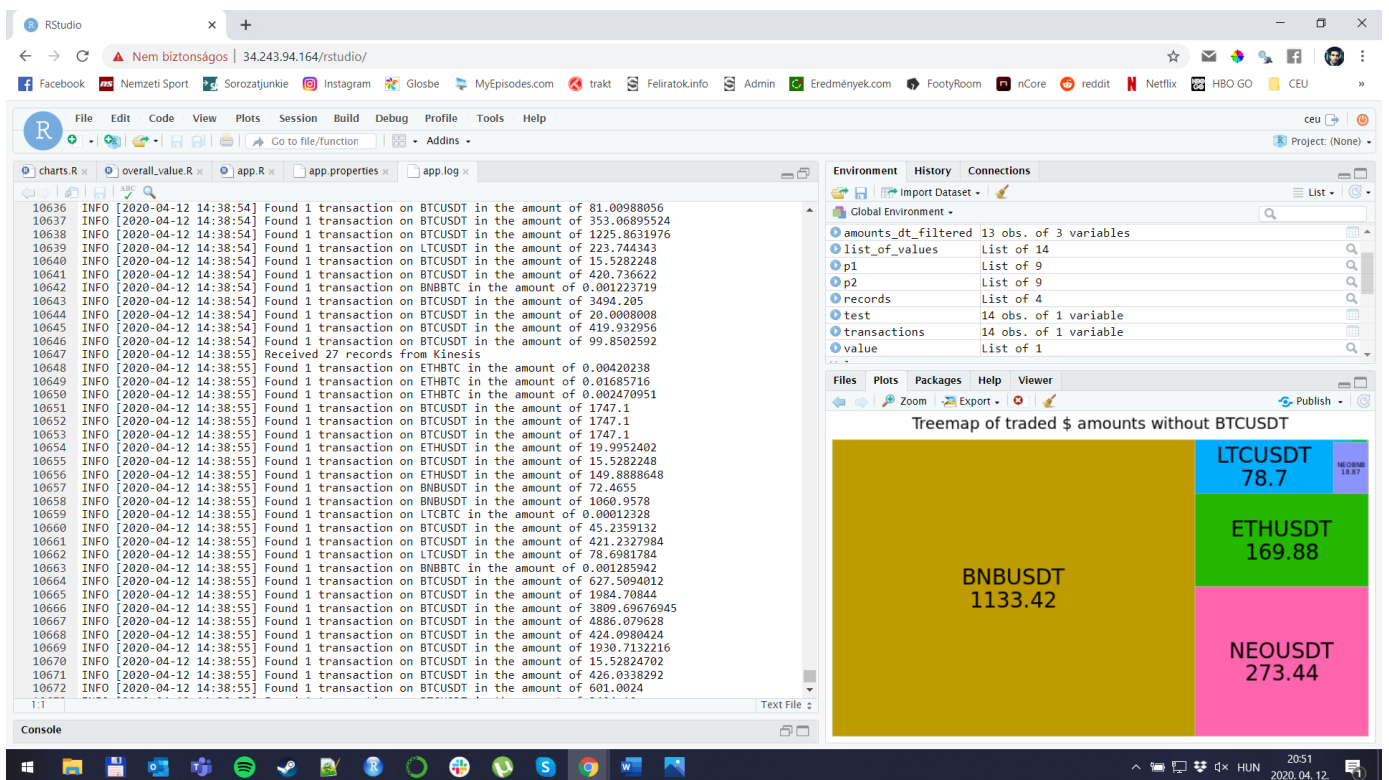
I created `app.R` based on the file from the lecture with the following modifications:

- Since the task was to show the overall amount of coins exchanged on Binance per symbol in the most recent micro-batch, in addition to the `symbol`, I also had to extract the `price` (named `p` in the stream) and the `quantity` (`q`) for each transaction so that I could calculate `amount` by multiplying the two.
- I'm initiating an empty data table (`processed_data`) before the for-loop for the records in the batch.
- I'm appending each record's `symbol` and `amount` to this data table.
- After processing the records from the loop, I'm aggregating by `symbol` to get the overall traded amount for each symbol in the batch.
- Only after these steps am I calling `redisIncr` to store the aggregated values in Redis.



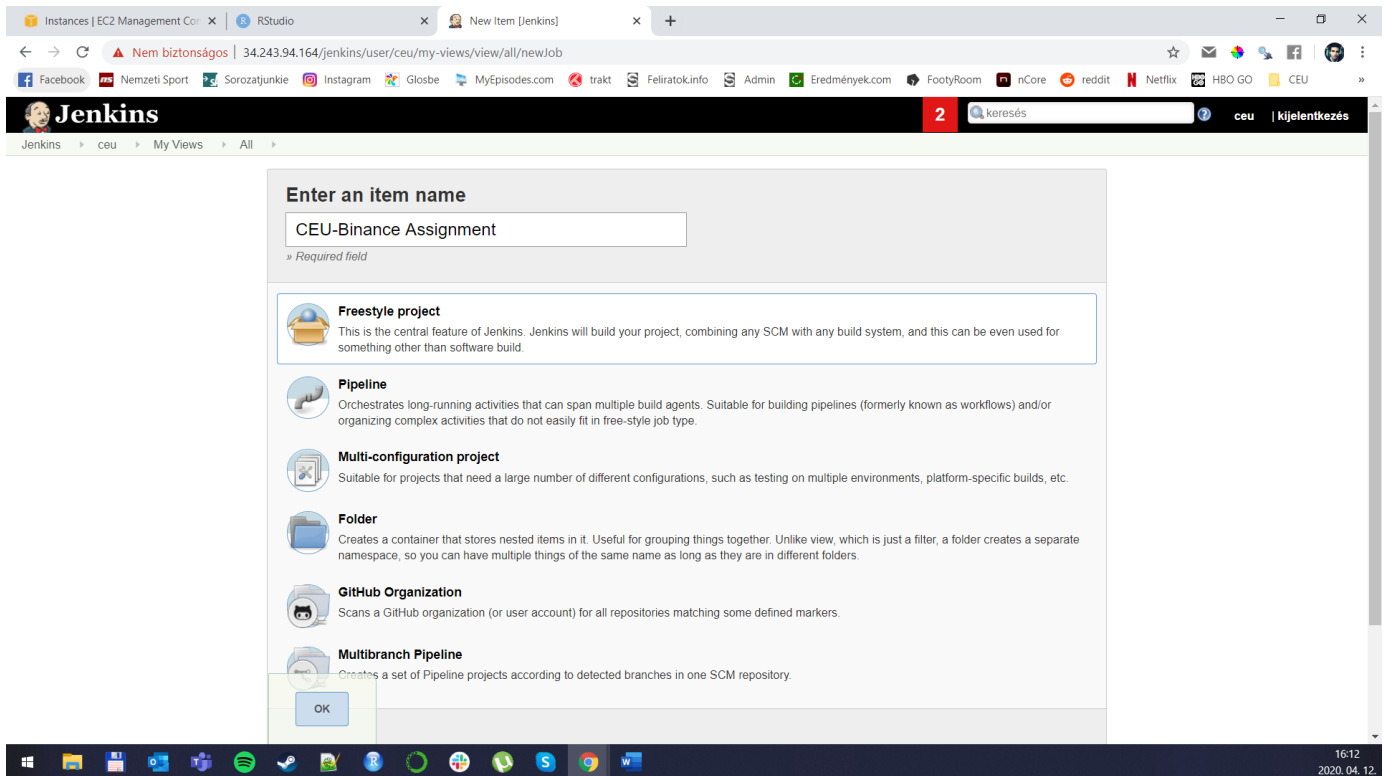


app.log logs all incoming batches with additional information about the received data:

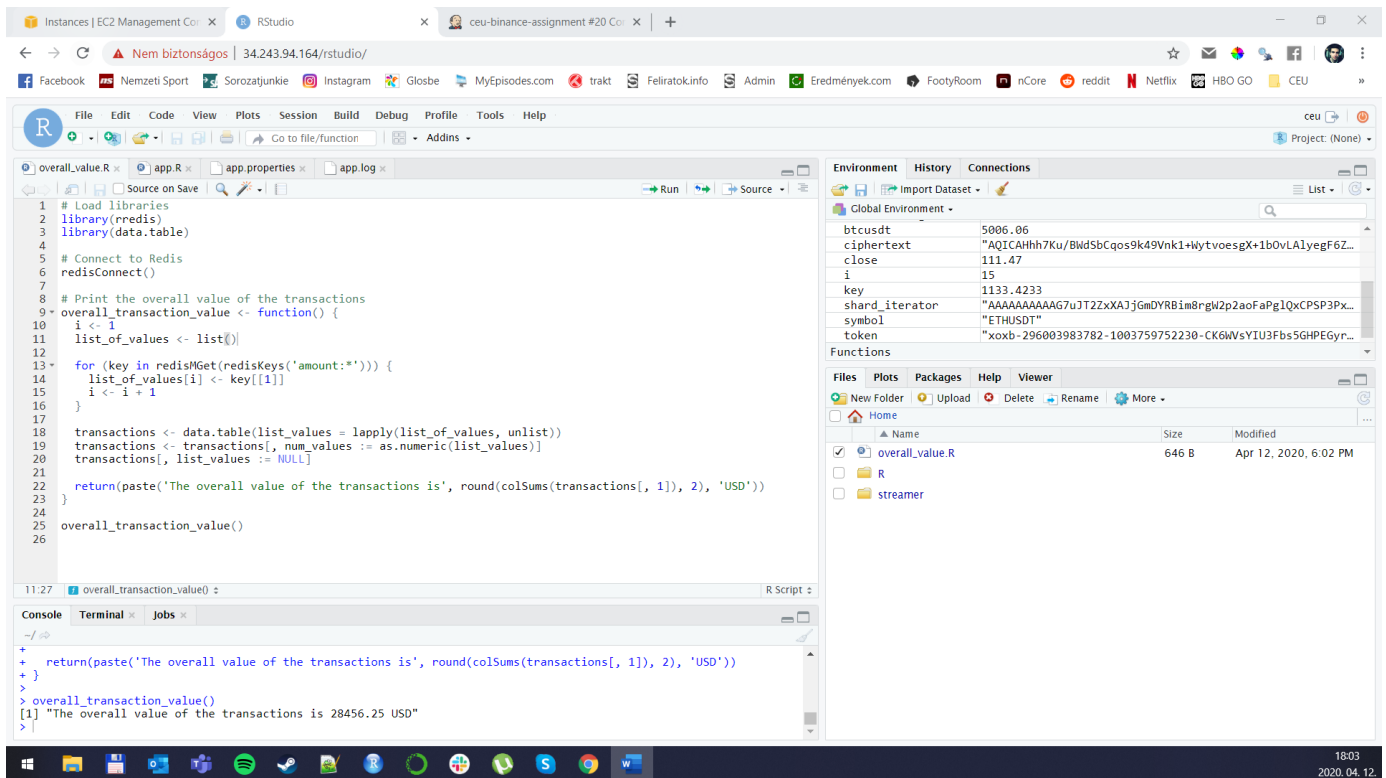


Jenkins

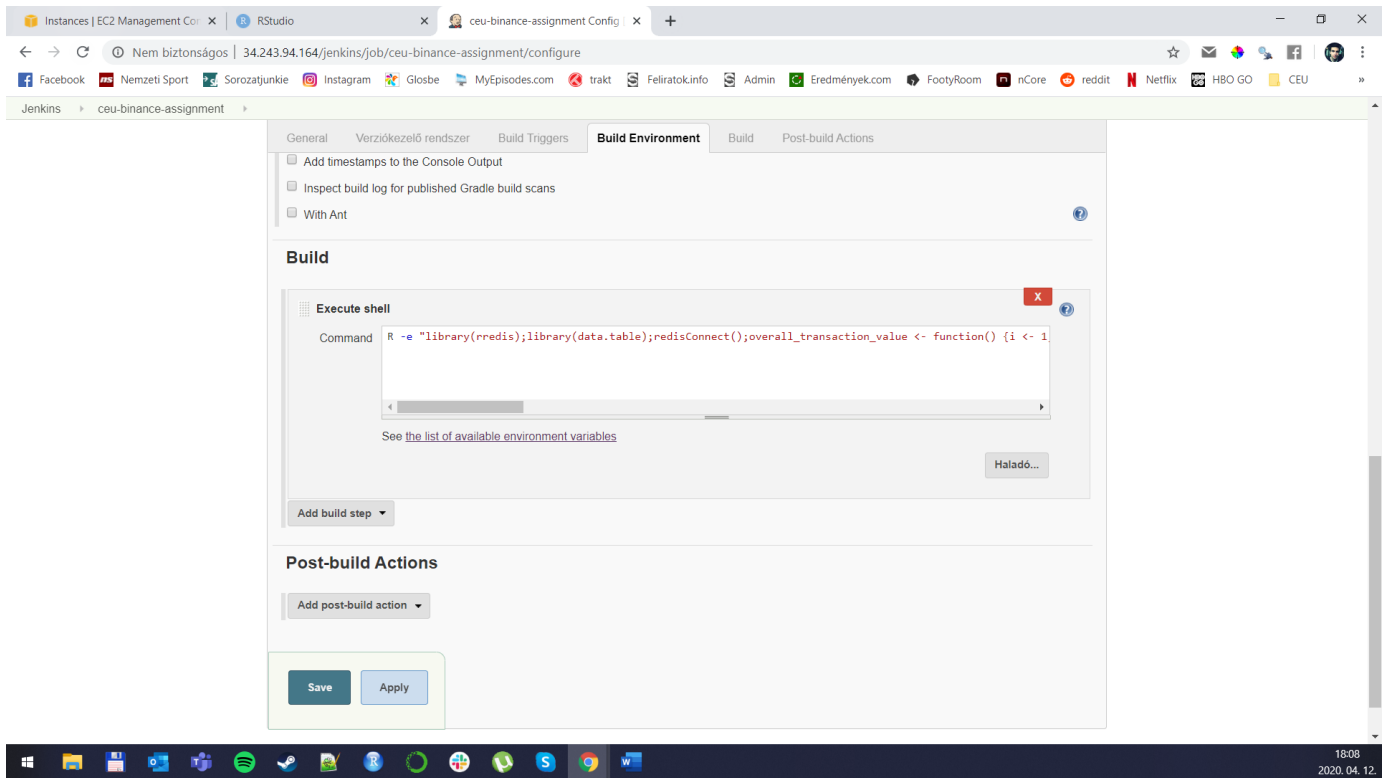
I created a new Jenkins job as a Freestyle project under the name `CEU-Binance Assignment` that I later renamed to `ceu-binance-assignment` to keep consistency with the naming convention I applied so far.



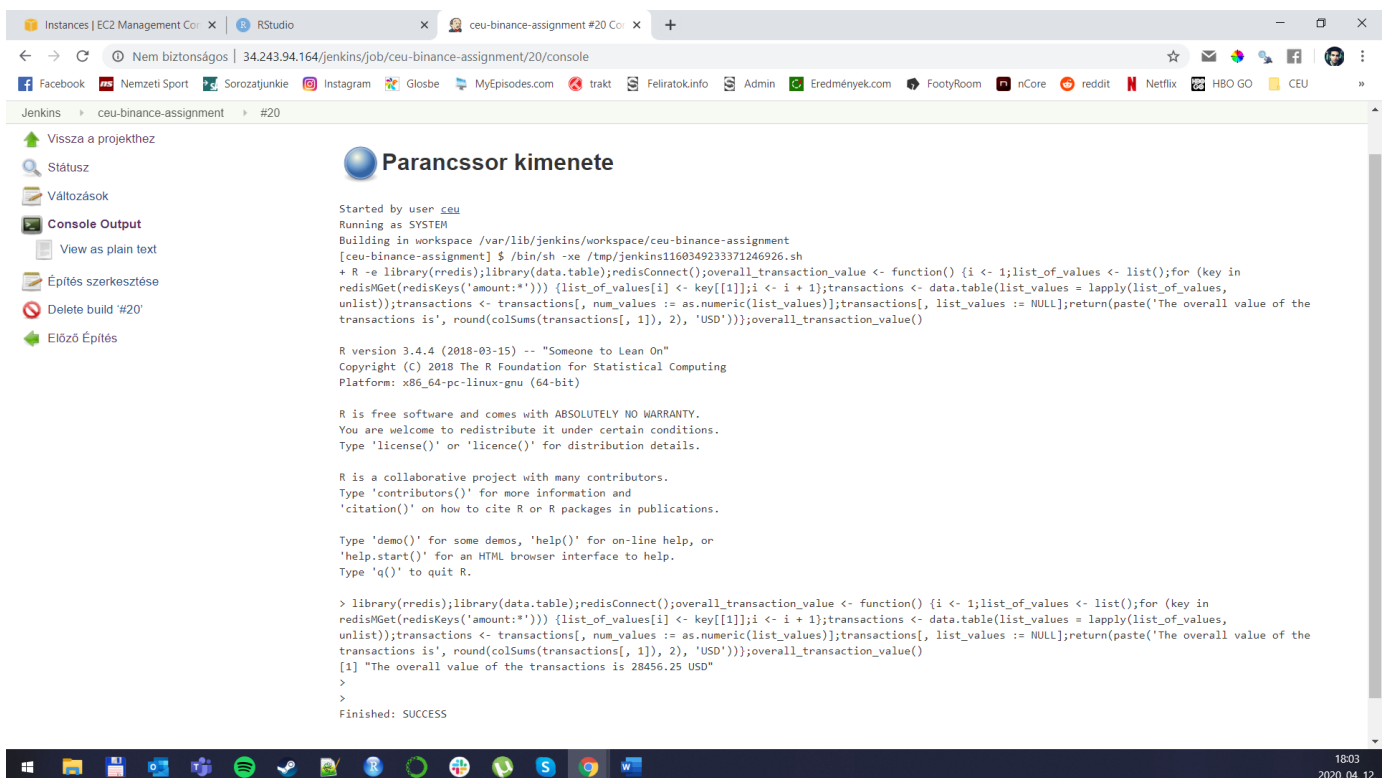
The script being called from Jenkins is the `overall_value.R` file which connects to Redis and creates a list from the key-value pairs that are stored in the Redis cache. As these items are coming through as a list, I have to unnest the values with `lapply`, and then cast them as `numeric`. After dropping the original list, I'm able to aggregate the column and calculate the sum of the transactions in the last batch. The `overall_transaction_value` function takes care of the data processing and transformation, and finally prints out the overall value of the transaction in USD.



Unfortunately I kept running into a “file not found” issue in Jenkins when trying to run `Rscript /home/overall_value.R`, so I ended up pasting the whole code snippet into “Execute shell command”:



The job ran successfully (after a couple of tries, this was run #20), and returned the overall value for the last batch:

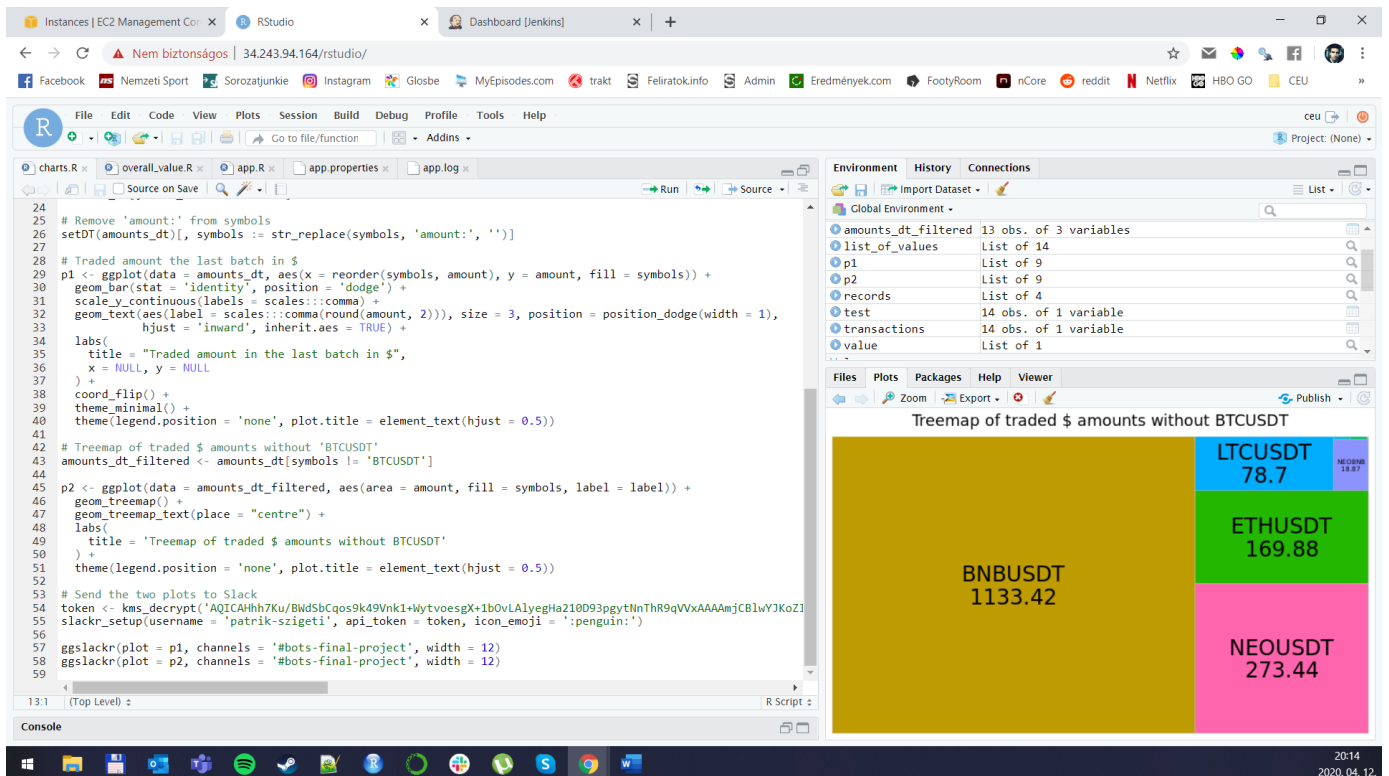
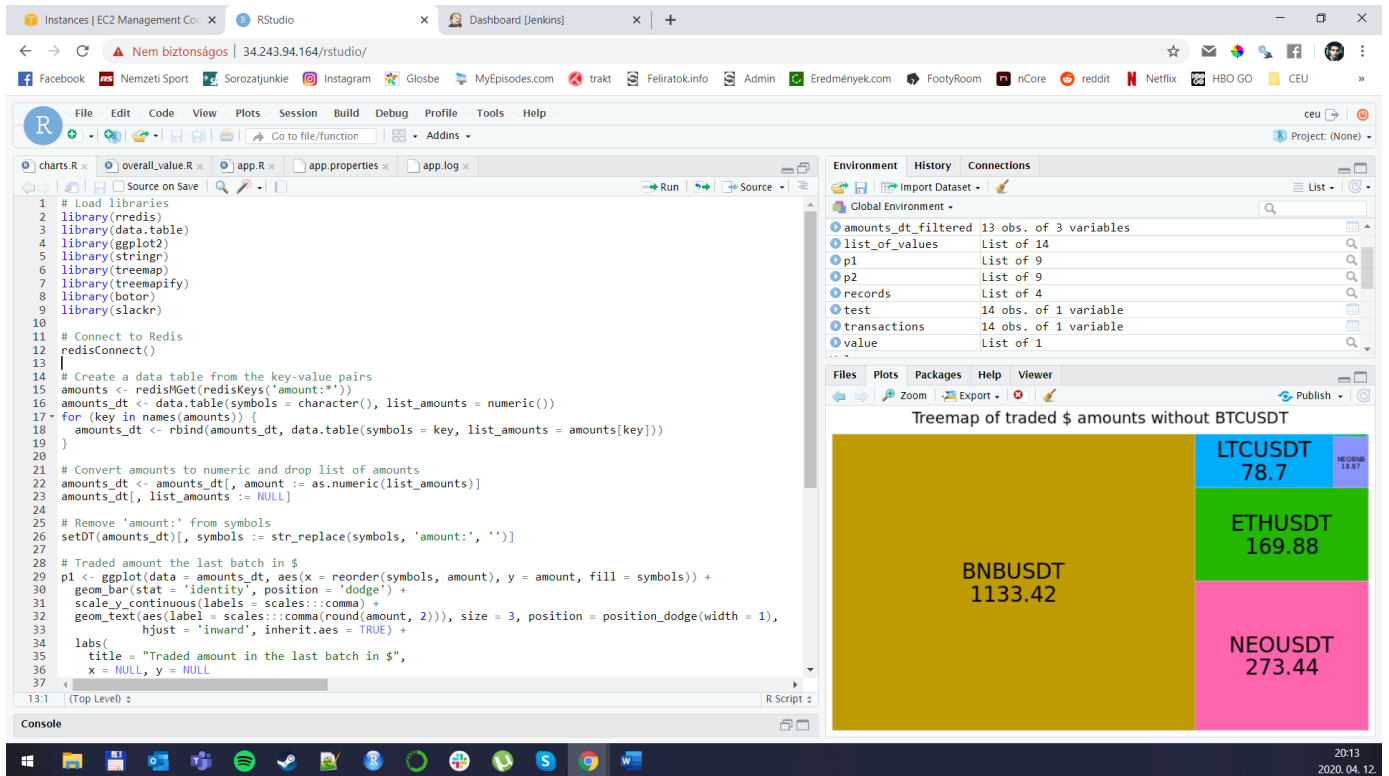


Plotting and Slack

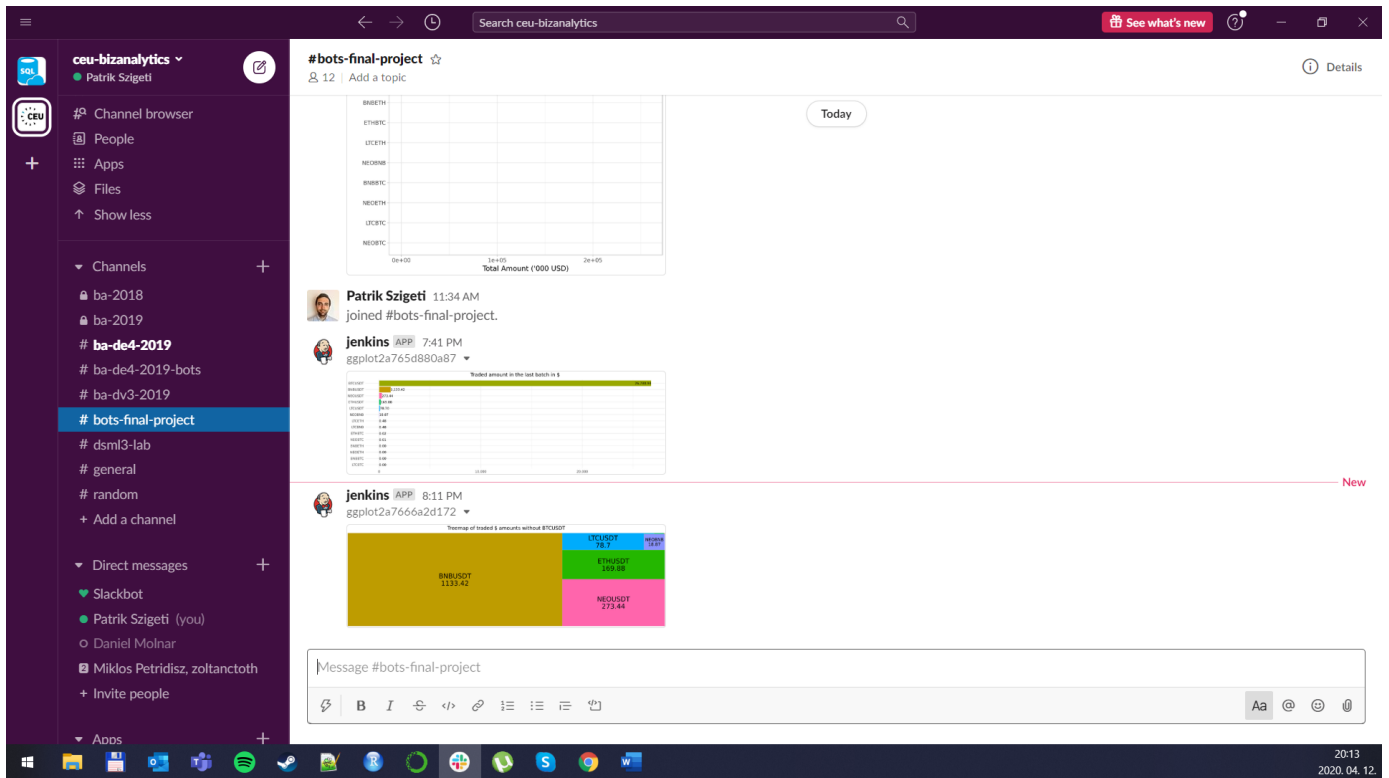
I created two plots to post to #bots-final-project, but before that, I had to create a data table from the key-value pairs, and then transform the amounts that were coming through as a list to numeric values. I also cleaned up the symbols column a bit, and removed the amount: prefix that I used to store the values in Redis.

- My first plot is a bar chart showing the traded amount in the last batch in USD.

- The second plot shows a treemap of the traded amounts without the amount for BTCUSDT, since otherwise that would take up a significant area from the chart, and I wanted to look at the “smaller guys”. For this, I also had to install the `treemapify` library to the server.



I posted both charts to the Slack channel:



Stop the instance

And I didn't forget to stop my instance either:

