Deployment of the Sentieon software to AWS

Release 202112.07

Sentieon, Inc

Apr 24, 2023

Contents

1 Introduction

1.1 Overview of the software

The Sentieon® software provides fast and efficient data processing for genomic data. Supported applications include secondary analysis of sequence data from whole-genome, whole-exome, targeted, and RNA sequence data for alignment, preprocessing, germline and somatic variant calling.

The Sentieon® Genomics software enables rapid and accurate analysis of next-generation sequence data. The Sentieon® DNAseq pipelines enable germline variant calling on hundreds of thousands of samples simultaneously. The Sentieon® TNseq pipelines enable accurate calling of somatic variants in paired tumor-normal samples or in an unpaired tumor samples. The Sentieon® Genomics software produces more accurate results than other tools on third-party benchmarks with results obtained in one tenth the time of comparable pipelines.

1.2 Platform requirements

Sentieon® Genomics software is designed to run on Linux and other POSIX-compatible platforms.

For Linux systems, we recommended using the following Linux distributions or higher: RedHat/CentOS 6.5, Debian 7.7, OpenSUSE-13.2, or Ubuntu-14.04.

Sentieon® Genomics software requires at least 16GB of memory, although we recommend using 64 GB of memory. The alignment step is typically the most memory consuming stage in most pipelines.

Sentieon® Genomics software processes file data at speeds around 20-30MB/s per core using a state of the art server. In order to take advantage of the maximum speed that the software can provide, we recommend that you use high-speed SSD hard drives in your system, preferably two identical drives in a high speed RAID 0 striped volume configuration. When using a RAID 0 configuration, it is advised to use the drives to only store intermediate files, and move out any final results or important information out of those hard drives; this way, your server should have additional storage space to store the results.

The required hard disk space is pipeline and data specific and can vary significantly between use-cases. A general rule of thumb is that the software can use approximately 3x the input file size for intermediate file storage during data processing.

2 Multi-Availability Zone Deployment on AWS

The Sentieon® software is controlled by a license and requires an active Sentieon® license server to run. Deployment on AWS requires setup and configuration of the Sentieon® license server along with setup and configuration of one or more instances for genomic data processing. A minimum deployment will use the following resources and should take approximately 20 minutes:

- An AWS VPC. The default VPC configuration with the default subnet(s), internet gateway, and route table is assumed in this deployment guide.
- Security groups that can be used to run the Sentieon® license server and the compute nodes.
- A (t2.nano) instance running the Sentieon® license server.

The Sentieon® software can be deployed onto additional instances for additional data processing capacity.

With a minimum deployment, customers will be charged by AWS EC2 for the instance running the license server and for outbound communication from the Sentieon® license server to the Sentieon® master license server. Adding additional comptute instances to the deployment will increase costs by the price of those instances. Data processing may incur additional costs for data transfer, etc. The Sentieon® software is a proprietary software and this deployment requires a license for the Sentieon® software.

This deployment is supported in the following AWS Regions:

- US East (Ohio)
- US East (N. Virginia)
- US West (N. California)
- US West (Oregon)
- Africa (Cape Town)
- Asia Pacific (Hong Kong)
- Asia Pacific (Jakarta)
- Asia Pacific (Mumbai)
- Asia Pacific (Osaka)
- Asia Pacific (Seoul)
- Asia Pacific (Singapore)
- Asia Pacific (Sydney)
- Asia Pacific (Tokyo)
- Canada (Central)
- Europe (Frankfurt)
- Europe (Ireland)
- Europe (London)

- Europe (Milan)
- Europe (Paris)
- Europe (Stockholm)
- Middle East (Bahrain)
- South America (São Paulo)

This deployment assumes basic familiarity with the VPC and EC2 services on AWS and familiarity with the Linux command-line interface.
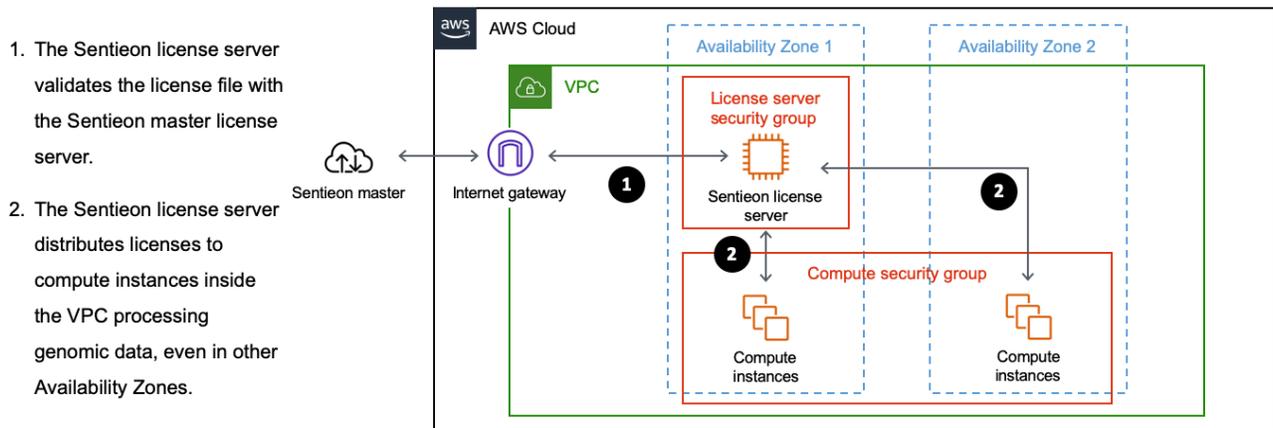
Multi-Availability Zone Deployment on AWS

Fig. 2.1: Overview of the Sentieon deployment architechure on AWS

2.1 Deployment of the Sentieon® license server

In order to use the software on AWS, you will need to follow the following installation instructions to deploy a Sentieon® License server to your VPC.

1. Choose an AWS Region to work in. Choose (or create) the VPC where you will run the Sentieon® tools.
2. Find and record the CIRD block for your VPC.
3. Create a security group for the license server (Fig. 2.2 and Fig. 2.3). The security group must:

- Allow inbound TCP communication at a specific port (we use 8990 by default as it is not typically used by other applications). We highly recommend that users choose a port above 1024 so the license server can be run as a non-root user. This rule is used to accept inbound communication from the compute nodes to the license server. You should open TCP at the desired port across your VPC's CIDR block (172.31.0.0/16 in Fig. 2.2) to whitelist traffic from within your VPC.
- Allow outbound HTTPS communication to Sentieon® license master at master.sentieon.com (IP 52.89.132.242). This is necessary for license validation.
- Allow inbound SSH. This is necessary for administration.
- (Recommended) Allow ICMP communication. This allows for PMTU discovery.

4. Launch a t2.nano instance to run the license server in the selected VPC. Be sure to assign the instance to the security group created in step 3.
5. Send your Sentieon® support representative the Private IP address of your instance, together with the TCP port you opened in step 3.
6. Download the Sentieon® tools and your license file to the instance.
7. On the t2.nano instance, start the license server with the following command. Note that the license server does not need to be started as a root user if a port above 1024 is chosen in step 3.

```
sentieon licsrvr --start [-l <licsrvr_log>] <license_file>
```

8. (Optional) Confirm the license server is working correctly and serving licenses to the license server instance by running following commands. The second command will return the number of available licenses.

```
sentieon licclnt ping -s <IP_addresss>:<PORT> || (echo "Ping Failed"; exit 1)
sentieon licclnt query -s <IP_address>:<PORT> klib
```

9. Create a security group for the compute nodes (Fig. 2.4 and Fig. 2.5). The security group should:

- Allow inbound SSH. This is necessary for administration.
- Allow outbound TCP communication at the port chosen in step 3, which will be open across your VPC's CIDR block.
- (Recommended) Allow ICMP communication to facilitate communication between the license server and compute nodes.

10. Launch a t2.nano instance within the same VPC to test the license server. Be sure to assign the instance to the security group created in step 9.
11. Download the Sentieon® software to the newly launched instance and confirm the license server is working and serving licenses within the VPC by running the following commands on the instance. The second command will return the number of available licenses.

```
sentieon licclnt ping -s <IP_addresss>:<PORT> || (echo "Ping Failed"; exit 1)
sentieon licclnt query -s <IP_address>:<PORT> klib
```
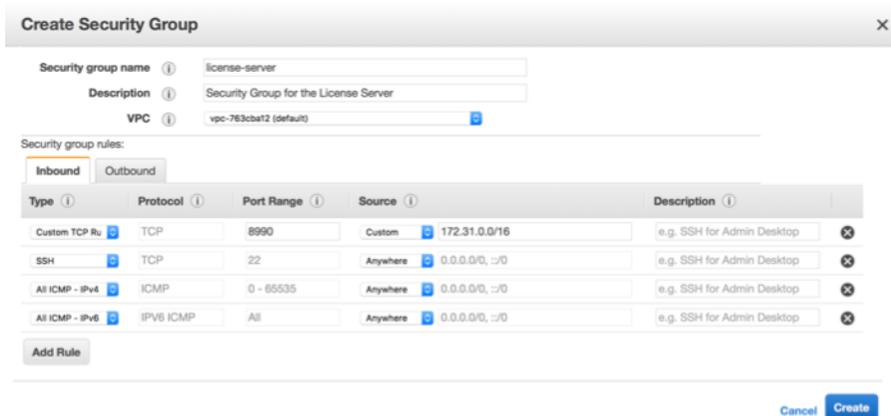

Fig. 2.2: Example license server inbound security group rules

After the license server has been deployed, the license server can be monitored by running the `sentieon licclnt ping` and `sentieon licclnt query` commands on a separate instance, as described in steps 10 and 11, above.
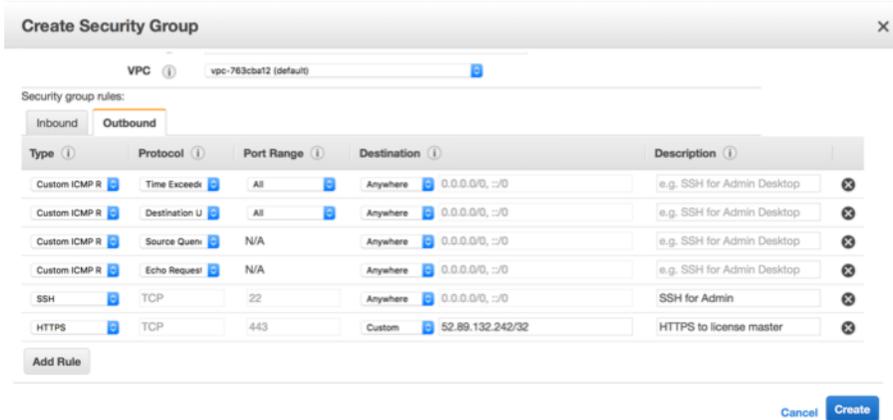
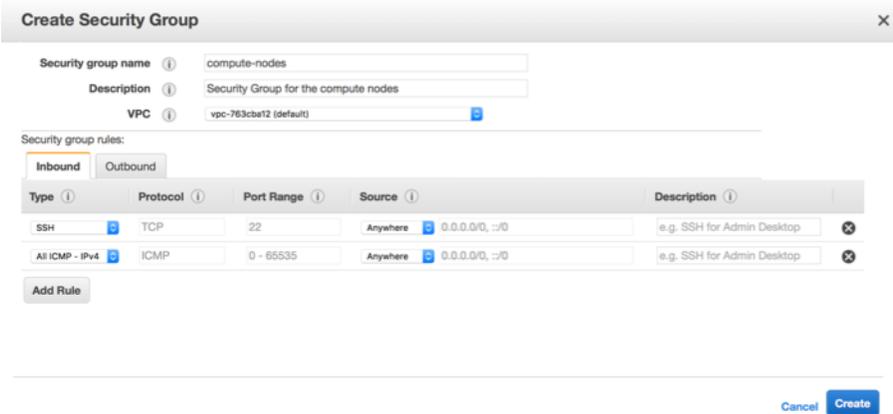Fig. 2.3: Example license server outbound security group rules

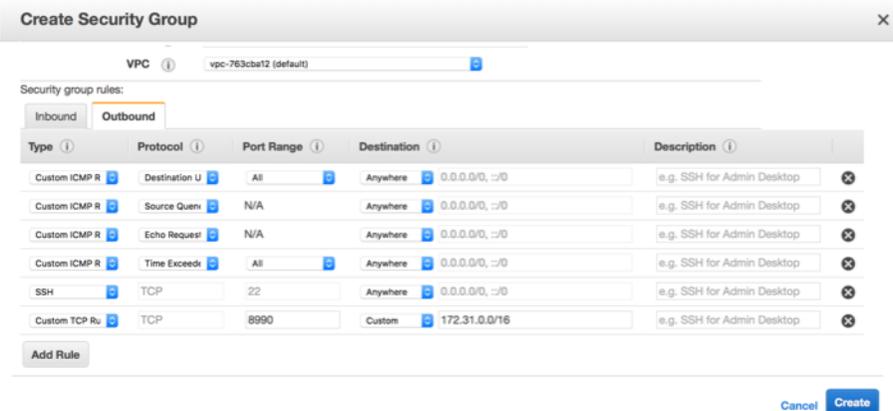Fig. 2.4: Example compute nodes inbound security group rule

Fig. 2.5: Example compute nodes outbound security group rules

2.2 Deployment of one or more instances for genomic data processing

After the license server is deployed in your VPC, you can deploy the software to one or more additional instances for genomic data processing. These instances can be deployed to any Availability Zone in the VPC. The following steps provide guidance for deploying the Sentieon® software onto newly launched EC2 instances:

1. Start an EC2 instance that meets the platform requirements for the Sentieon® software inside your VPC. Assign the instance to the security group created in *step 9* in the above section on deployment of the license server..
2. Download the Sentieon® tools to the newly launched instance.
3. Set an environment variable in your system to indicate the location of the license server.

```
export SENTIEON_LICENSE=<LICSRVR_INTERNAL_IP>:<LICSRVR_PORT>
```

The new instance is now ready to process data and data processing does not require root privileges. Please see the Sentieon® software manual at, https://support.sentieon.com/manual/ for more information on the functionality included in the Sentieon® software.

2.3 Upgrading the deployment and restarting the license server

Most new releases of the Sentieon® software will not affect the function of the license server. In the event of a software update affecting the license server, the following steps can be used to update the license server or may be used to restart the license server to recover the license server.

1. Download the updated release of the Sentieon® tools to the instance.
2. Run the following command on the license server instance to stop the license server.

```
sentieon licsrvr --stop <license_file>
```

3. Using the updated software package, run the following command to restart the license server.

```
sentieon licsrvr --start [-l <licsrvr_log>] <license_file>
```

3 Troubleshooting

If you encounter issues with the software, please refer to the Sentieon Software Manual[1] and Quick Start Guide[2]. You may also contact the Sentieon® support team at support@sentieon.com.

---

https://support.sentieon.com/manual
https://support.sentieon.com/quick_start