ThoughtSpot on AWS Quick Start Guide

Version 4.2

February 2017

Contents

Chapter 1: Welcome to ThoughtSpot

Topics:

Congratulations on purchasing the ThoughtSpot instance. This guide will get you started with setting up the Amazon Web Services (AWS) virtual machine (VM) offering.

We hope your experience with ThoughtSpot is excellent. Please let us know how it goes, and what we can do to make it better.

Contact ThoughtSpot

You can contact ThoughtSpot by phone, mail, email, or by filing a support ticket.

File a support ticket

If you encounter a technical issue, file a support ticket using the Support Portal ticket filing system at:

http://support.thoughtspot.com/

Please provide as much detail as possible about your issue, to help us resolve it quickly.

You need a Support Portal login to file a ticket. Please contact ThoughtSpot to get an account, if necessary.

Address

ThoughtSpot, Inc.

1 Palo Alto Square, Building 1, Suite 200

Palo Alto, CA 94306

Phone numbers

Table 1: Phone numbers

Phone Number	Description
1-800-508-7008 ext 1	ThoughtSpot Support
1-800-508-7008	Toll free number for ThoughtSpot headquarters.

Email

Table 2: Email addresses

Reason for contacting	Email
For sales inquiries.	sales@thoughtspot.com
For customer support and software update inquiries.	support@thoughtspot.com
For other inquiries.	hello@thoughtspot.com

Chapter 2: Introduction

Topics:

Before you set up your ThoughtSpot instance, here is some information about the Amazon Web Services (AWS) cloud, Amazon Elastic Compute Cloud (EC2), Amazon Elastic Block Store (EBS), and how they all work with ThoughtSpot.

About AWS

AWS is a secure cloud services platform offered by Amazon. Using ThoughtSpot on AWS allows you to easily add instances as your dataset grows.

You can do everything you'd normally want to do in a traditional database center with AWS. It features an on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing.

Amazon EC2 is based on instance types and the region in which your instances are running. When you are connected to an instance, you can use it just like you use any other server. There is no minimum fee and you only pay for what you use.

Using Amazon EC2 lets you develop and deploy applications faster since there is no need to manage hardware. Therefore, it is easy to scale and manage computing capacity.

As persistent block level storage volumes, Amazon EBS helps with scaling your EC2 instances. Each EBS volume is automatically replicated to protect you from component failure, and offers low-latency performance.

About ThoughtSpot on AWS

AWS can provide lots of memory and CPU for your ThoughtSpot instance, and it can be easily updated from development instances.

Your database capacity will determine the number of instances you'll need and the instance network/storage requirements. In addition, you can go with multiple VMs based on your dataset size.

The security group setting of your ThoughtSpot instance on AWS is up to you. You can find more information about which ports to open in the network ports reference.

Chapter 3: Configuration

Topics:

ThoughtSpot engineering has performed extensive testing of the ThoughtSpot appliance on various Amazon Elastic Compute Cloud (EC2) and Amazon Elastic Block Store (EBS) configurations for best performance, load balancing, scalability, and reliability.

Configuration options

You can find information here on which configuration of memory, CPU, storage, and networking capacity you should be running for your instances. There are also details on how to configure your placement groups.

Hardware configurations

There is only one available hardware configuration for deploying ThoughtSpot on Amazon:

- r4.16xlarge

Below are charts depicting the specifications for the configuration for EC2 and EBS requirements.

Table 3: EC2 requirements for deploying on Amazon

Instance name	Data capacity	vCPUs	DRAM
r4.16xlarge	Up to 250 GB	64	488 GB

Table 4: EBS requirements for deploying on Amazon

Instance name	Data capacity	Root volume (SSD)	Data volume (SSD or HDD)
r4.16xlarge	Up to 250 GB	1 vol 200 GB	2 vols 400 GB each

Attention: Both EC2 and EBS requirements must be fulfilled to deploy on Amazon.

ThoughtSpot software license sizes

ThoughtSpot only sells software licenses in multiples of 250 GB of data. So you can start with 250 GB, and add increments of 250 GB each time your data capacity needs increase. You can also choose to start off with more than 250 GB of data, as long as you know the best fit configuration for your data volume.

Lego blocks

If you aren't sure what kind of configuration you need, it might help to think of the hardware configurations in terms of simple Lego blocks. The r4.16xlarge size can be seen as a 250 GB block.

Note: ThoughtSpot does not support sizes other than r4.16xlarge.

Since the minimum data volume offered is 250 GB, you would need one r4.16xlarge block to match the data capacity. This scales linearly. So, 500 GB would require two r4.16xlarge blocks.

Placement groups

A placement group is a logical grouping of instances within a single availability zone. Placement groups are recommended for applications that benefit from low network latency, high network throughput, or both.

ThoughtSpot relies on high connectivity between nodes of a cluster, which is why creating a placement group is recommended. Being in same placement group will give you the best shot at the highest bandwidth across AWS EC2 instances and the lowest latencies. This will make the node-node network reach the closest AWS promised specs. Our default recommendation for a multi-instance setup requires a placement group since it works best for our application performance. Also, AWS will provide jumbo frames (9000 MTU) support in such situations, and they don't charge extra for being in the same placement group. Having said that, ThoughtSpot will still work with EC2s in the cluster across placement groups in an availability zone.

Related information:

EC2 instance types

EC2 pricing

EBS pricing

Placement groups

Chapter 4: Installation and setup

Topics:

Here is an overview of the installation and setup information for getting ThoughtSpot on Amazon Web Services (AWS) Elastic Compute Cloud (EC2).

After you've determined your configuration options, you must setup your virtual machines (VMs) using an Amazon Machine Image (AMI). This AMI will be shared with you by ThoughtSpot.

About the AMI

An AMI is a preconfigured template that provides the information required to launch an instance. Check with your ThoughtSpot contact to learn about the latest version of the ThoughtSpot AMI.

You must specify an AMI when you launch an instance. An AMI includes the following:

- A template for the root volume for the instance (for example, an operating system, an appliance server, and applications).
- Launch permissions that control which AWS accounts can use the AMI to launch instances.
- A block device mapping that specifics the volumes to attach to the instance when it's launch.

About ThoughtSpot's AMI

The ThoughtSpot AMI comes provisioned with the custom ThoughtSpot image to make hosting simple. Once you've provided your AWS account ID and region where the VMs will be hosted, ThoughtSpot will share the current ThoughtSpot base AMI with you. This AMI has ThoughtSpot specific applications on an Ubuntu 12.04 base image. The EBS volumes required for ThoughtSpot install in AWS comes as part of the AMI. When you launch an EC2 instance from this image, the EBS volumes automatically get sized and provisioned. The storage attached to the base AMI is 200 GB (xvda), 2X400 GB (xvdb), and SSD gp2. It contains the max disks so that it can take care of the full load of the VM.

Launch an instance

Follow these steps to set up the VMs and launch ThoughtSpot.

ThoughtSpot instances on AWS need AWS EC2 instances to be provisioned in the AWS account before ThoughtSpot can be installed and launched. Please make sure you follow the guidelines below for your EC2 details:

- EC2 instance type: r4.16xlarge.
- Networking requirement: 10GbE network is needed between the VMs. This is the default for the chosen VM type.
- Security: VMs need to be accessible from each other, which means they need to be on the same Amazon Virtual Private Cloud (VPC) and subnetwork. Additional external access may be required to bring data in/out of the VMs to your network.
- Number of EC2 instances needed: Based on the datasets, the number of EC2 instances needed will vary. Also for staging larger datasets (> 50 GB per VM), there may be a need to provision additional attached EBS volumes that are SSD gp2 provisioned.

To set up the VMs and launch ThoughtSpot:

1. Log in to your AWS account from the AWS Amazon sign in page.
2. Provide ThoughtSpot Support with your AWS account ID and the region where the VMs will be hosted. They will then grant you permissions and share the current ThoughtSpot base AMI with you.

   Tip: You can find your account ID and region on the top right corner of the AWS console.

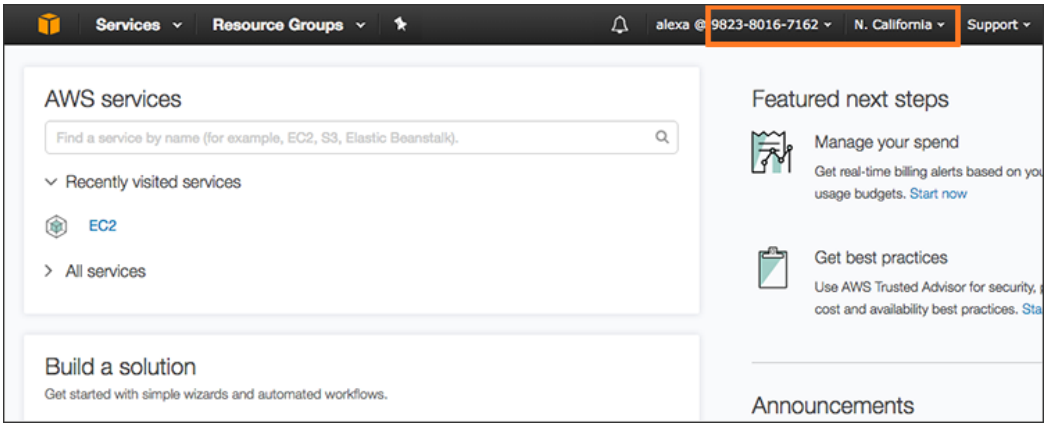

Figure 1: AWS account ID and region

3. Navigate to the EC2 service dashboard by clicking **Services**, then select **EC2**.

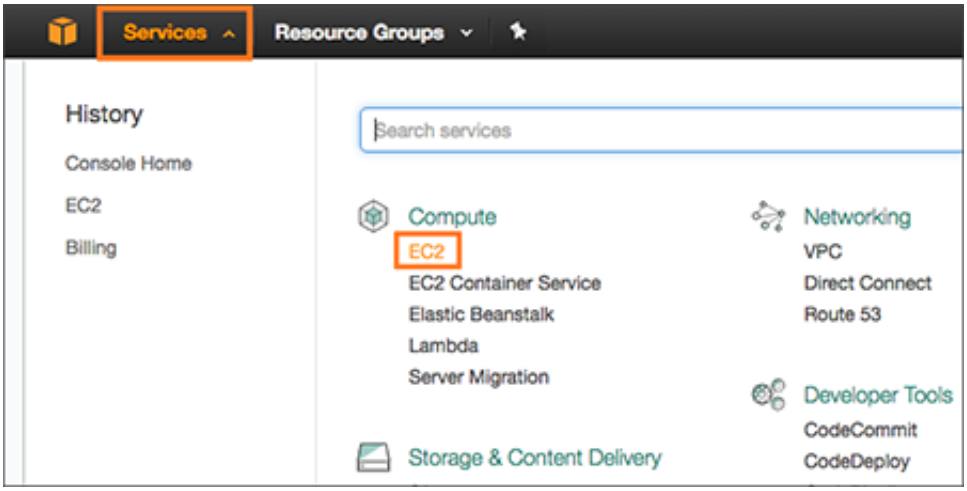

Figure 2: Navigate to the EC2 Dashboard

4. Make sure your selected region is correct on the top right corner of the dashboard. If not, select a different region you would like to launch your instance in. Let ThoughtSpot Support know if you change your region.

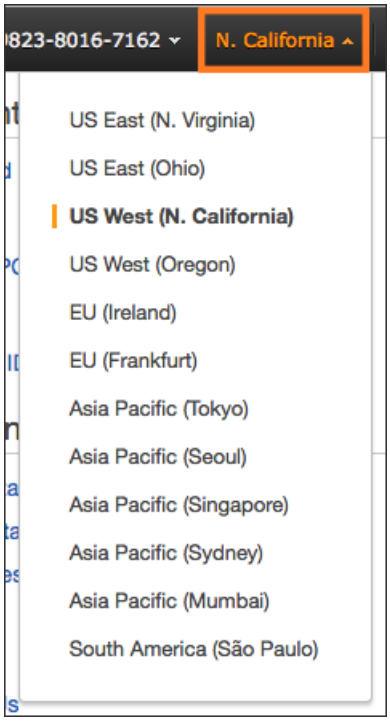

Figure 3: Select a region to launch your instance in

5. Create an instance by clicking **Launch Instance**.

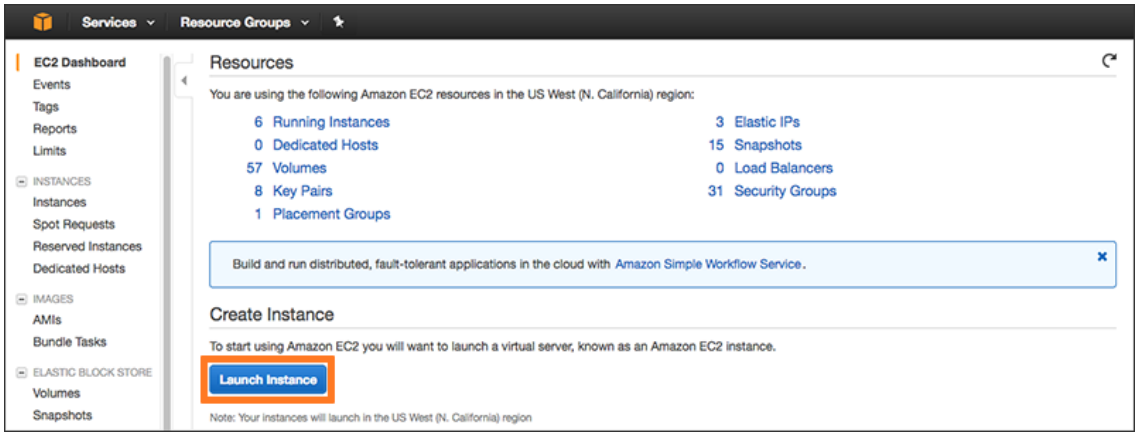

Figure 4: Launch an instance

6. Select the appropriate AMI from the AMI Selection step by clicking **Select**. The ThoughtSpot shared AMI can be found under the **My AMIs** tab.

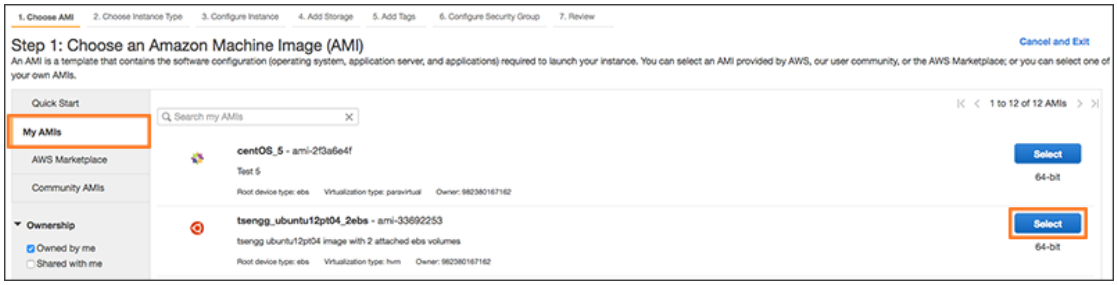

Figure 5: Select the AMI

7. Select r4.16xlarge as the instance type. Then click **Next: Configure Instance Details**.
8. Configure the instances by choosing the number of EC2 instances you need based on your EC2 details. The instances need to be on the same VPC and subnetwork. Then click **Next: Add Storage**.
9. The default storage specified by the ThoughtSpot AMI should be populated. Optionally, you can add extra storage. Based on the dataset size requirement you might need to provision and prepare (formatting/file system placement) an extra storage of 400 GB per VM that is SSD gp2 provisioned. Click **Next: Add Tags** when you are done modifying the storage size.
10. Set a name for tagging your instances. Then click **Next: Configure Security Group**.
11. Select an existing security group to attach new security groups to such that it meets the security requirements for ThoughtSpot.

    Note: Security setting for ThoughtSpot:

    - The VMs need intragroup security, i.e. every VM in a cluster needs to be accessible from one another. For easier configuration, it is better to open all accesses from across VMs in a cluster.
    - Additionally, more ports need to be opened on the VM to provide data staging capabilities to your network. Check the network ports

reference to determine the minimum required ports that need to be opened for your ThoughtSpot appliance.

12. Click **Review and Launch**. After you have reviewed your instance launch details, click **Launch**.
13. Choose a key pair. A key pair consists of a public and private key used to encrypt and decrypt login information. If you don't have a key pair, you should create one, otherwise you won't be able to SSH into the AWS instance later on.
14. Click **Launch Instances**. Wait a few minutes for it to fully start up. Once it has started up, it will show up on the EC2 console.
15. Contact ThoughtSpot Support to complete your ThoughtSpot installation.

Chapter 5: ThoughtSpot on AWS reference

Topics:

This section contains a reference for ThoughtSpot on Amazon Web Services (AWS).

Network ports

This reference lists the potential ports to open when setting up your security group.

Network ports

For regular operations and for debugging, there are some ports you will need to keep open to network traffic from end users. Another, larger list of ports must be kept open for network traffic between the nodes in the cluster.

Required ports for operations and debugging

The following ports need to be opened up to requests from your user population. There are two main categories: operations and debugging.

Table 5: Network ports to open for operations

Port	Protocol	Service Name	Direction	Source	Destination	Description
22	SSH	SSH	bidirectional	Administrators IP addresses	All nodes	Secure shell access. Also used for scp (secure copy).
80	HTTP	HTTP	bidirectional	All users IP addresses	All nodes	Hypertext Transfer Protocol for website traffic.
443	HTTPS	HTTPS	bidirectional	All users IP addresses	All nodes	Secure HTTP.
12345	TCP	Simba	bidirectional	Administrators IP addresses	All nodes	Port used by ODBC and JDBC drivers when connecting to ThoughtSpot.

Table 6: Network ports to open for debugging

Port	Protocol	Service Name	Direction	Source	Destination	Description
2201	HTTP	Orion master HTTP	bidirectional	Administrator IP addresses	All nodes	Port used to debug the cluster manager.
2101	HTTP	Oreo HTTP	bidirectional	Administrator IP addresses	All nodes	Port used to debug the node daemon.

Port	Protocol	Service Name	Direction	Source	Destination	Description
4001	HTTP	Falcon worker HTTP	bidirectional	Administrator IP addresses	All nodes	Port used to debug the data cache.
4251	HTTP	Sage master HTTP	bidirectional	Administrator IP addresses	All nodes	Port used to debug the search engine.

Required ports for inter-cluster operation

Internally, ThoughtSpot uses static ports for communication between services in the cluster. Do not close these ports from inter-cluster network communications. In addition, a number of ports are dynamically assigned to services, which change between runs. The dynamic ports come from the range of Linux dynamically allocated ports (20K+).

Table 7: Network ports to open between the nodes in the cluster

Port	Protocol	Service Name	Direction	Source	Dest.	Description
80	TCP	nginx	inbound	All nodes	All nodes	Primary app HTTP port (nginx)
443	TCP	Secure nginx	inbound	All nodes	All nodes	Primary app HTTPS port (nginx)
2100	RPC	Oreo RPC port	bidirectional	All nodes	All nodes	Node daemon RPC
2101	HTTP	Oreo HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Node daemon HTTP
2181	RPC	Zookeeper servers listen on this port for client connections	bidirectional	All nodes	All nodes	Zookeeper servers listen on this port for client connections
2200	RPC	Orion master RPC port	bidirectional	All nodes	All nodes	Internal communication

Port	Protocol	Service Name	Direction	Source	Dest.	Description
						with the cluster manager
2201	HTTP	Orion master HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the cluster manager
2210	RPC	Cluster stats service RPC port	bidirectional	All nodes	All nodes	Internal communication with the stats collector
2211	HTTP	Cluster stats service HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the stats collector
2230	RPC	Callosum stats collector RPC port	bidirectional	All nodes	All nodes	Internal communication with the BI stats collector
2231	HTTP	Callosum stats collector HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the BI stats collector
2240	RPC	Alert manager	bidirectional	All nodes	All nodes	Port where alerting service receives alert events
2888	RPC	Ports used by Zookeeper servers for communication between themselves	bidirectional	All nodes	All nodes	Ports used by Zookeeper servers for communication between themselves
3888	RPC	Ports used by Zookeeper servers for communication	bidirectional	All nodes	All nodes	Ports used by Zookeeper servers for communication

Port	Protocol	Service Name	Direction	Source	Dest.	Description
		between themselves				between themselves
4000	RPC	Falcon worker RPC port	bidirectional	All nodes	All nodes	Port used by data cache for communication between themselves
4001	HTTP	Falcon worker HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the data cache
4021	RPC	Sage metadata service port (exported by Tomcat)	bidirectional	Admin IP addresses and all nodes	All nodes	Port where search service contacts metadata service for metadata
4201	HTTP	Sage auto complete server HTTP interface port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the search service
4231	HTTP	Sage index server HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the search service
4232	RPC	Sage index server metadata subscriber port	bidirectional	All nodes	All nodes	Port used for search service internal communication
4233	RPC	Sage index server RPC port	bidirectional	All nodes	All nodes	Port used for search service internal communication
4241	HTTP	Sage auto complete server HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Port used to debug the search service

Port	Protocol	Service Name	Direction	Source	Dest.	Description
4242	RPC	Sage auto complete server RPC port	bidirectional	All nodes	All nodes	Port used for search service internal communication
4243	RPC	Sage auto complete server metadata subscriber port	bidirectional	All nodes	All nodes	Port used for search internal communication
4251	RPC	Sage master RPC port	bidirectional	All nodes	All nodes	Port used for search service internal communication
4405	RPC	Diamond (graphite) port	bidirectional	All nodes	All nodes	Port used for communication with monitoring service
4500	RPC	Trace vault service RPC port	bidirectional	All nodes	All nodes	Trace collection for ThoughtSpot services
4501	HTTP	Trace vault service HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Debug trace collection
4851	RPC	Graphite manager RPC port	bidirectional	All nodes	All nodes	Communication with graphite manager
4852	HTTP	Graphite manager HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Debug graphite manager
4853	RPC	Elastic search stack (ELK) manager RPC port	bidirectional	All nodes	All nodes	Communication with log search service

Port	Protocol	Service Name	Direction	Source	Dest.	Description
4853	HTTP	Elastic search stack (ELK) manager HTTP port	bidirectional	Admin IP addresses and all nodes	All nodes	Debug log search service
5432	Postgres	Postgres database server port	bidirectional	All nodes	All nodes	Communication with Postgres database
8020	RPC	HDFS namenode server RPC port	bidirectional	All nodes	All nodes	Distributed file system (DFS) communication with clients
8080	HTTP	Tomcat	bidirectional	All nodes	All nodes	BI engine communication with clients
8787	HTTP	Periscope (UI) service HTTP port	bidirectional	All nodes	All nodes	Administration UI back end
8888	HTTP	HTTP proxy server (tinyproxy)	bidirectional	All nodes	All nodes	Reverse SSH tunnel
11211	Mem-cached	Memcached server port	bidirectional	All nodes	All nodes	BI engine cache
12345	ODBC	Simba server port	bidirectional	All nodes	All nodes	Port used for ETL (extract, transform, load)
50070	HTTP	HDFS namenode server HTTP port	bidirectional	All nodes	All nodes	Debug DFS metadata
50075	HTTP	HDFS datanode server HTTP port	bidirectional	All nodes	All nodes	Debug DFS data

Required ports for inbound and outbound cluster access

ThoughtSpot uses static ports for inbound and outbound access to a cluster.

Table 8: Network ports to open for inbound access

Port	Protocol	Service Name	Direction	Source	Dest.	Description
22	SCP	SSH	bidirectional	ThoughtSpot Support	All nodes	Secure shell access.
80	HTTP	HTTP	bidirectional	ThoughtSpot Support	All nodes	Hypertext Transfer Protocol for website traffic.
443	HTTPS	HTTPS	bidirectional	ThoughtSpot Support	All nodes	Secure HTTP.
12345	TCP	Simba	bidirectional	ThoughtSpot Support	All nodes	Port used by ODBC and JDBC drivers when connecting to ThoughtSpot.

Table 9: Network ports to open for outbound access

Port	Protocol	Service Name	Direction	Source	Destination	Description
443	HTTPS	HTTPS	outbound	All nodes	208.83. 110.20	For transferring files to thoughtspot.egnyte.com (IP address 208.83.110.20).
25 or 587	SMTP	SMTP or Secure SMTP	outbound	All nodes and SMTP relay (provided by customer)	All nodes	Allow outbound access for the IP address of whichever email relay server is in use. This is for sending alerts to ThoughtSpot Support.
389 or 636	TCP	LDAP or LDAPS	outbound	All nodes and LDAP server (provided by customer)	All nodes	Allow outbound access for the IP address of the LDAP server in use.

