Final Project Tutorial

Splunk

Group 9

Anvita Rayala

[Anvita.Rayala001@umb.edu](mailto:Anvita.Rayala001@umb.edu)

April 21, 2025,

A report submitted in partial fulfillment of

The Requirements for Big Data by **Professor Lei Zheng**

**Introduction:**

**What is Splunk?**

Splunk is a powerful data analytics platform used for collecting, indexing, and analyzing machine-generated data in real time. It excels at making sense of massive volumes of unstructured and semi-structured log data. Splunk helps organizations monitor systems, detect anomalies, generate reports, and respond quickly to security or operational issues. When deployed on cloud infrastructure like AWS, Splunk becomes even more scalable, enabling big data analysis across cloud-native services.

**Use cases:**

Splunk is particularly useful in scenarios involving large-scale infrastructure, high-volume log generation, or security monitoring. Some key use cases include:

* **Monitoring cloud-based systems and services (e.g., EC2, RDS, Lambda)**
* **Detecting anomalies and security breaches via CloudTrail or VPC Flow Logs**
* **Auditing user behavior across applications and infrastructure**
* **Real-time alerting and dashboard visualization.**

**Comparison with Alternatives:**

| **Technology** | **Key Strength** | **Deployment** |
| --- | --- | --- |
| **Splunk** | Real-time log indexing, alerts, dashboards | On-prem / AWS / Cloud |
| **ELK Stack** | Open-source and highly customizable | Self-managed |
| **AWS CloudWatch** | Native monitoring tool for AWS services | AWS-native |
| **Datadog** | Cloud-based, strong visualization and metrics | SaaS |
| **Graylog** | Good performance, open-source SIEM features | Self-managed |

**Advantages:**

**Real-time search and analytics on huge log datasets:**  
Splunk can ingest and analyze massive volumes of log data as it's generated, enabling instant visibility into systems and services.

**Intuitive dashboard and visualization tools:**  
It offer drag-and-drop dashboard creation with charts, maps, and gauges for easier monitoring and reporting.

**Supports both structured and unstructured data:**  
Splunk handles logs, events, metrics, and any machine data, whether it's cleanly formatted or raw.

**Scales easily with cloud deployment on AWS:**  
Using AWS, Splunk can be scaled vertically or horizontally depending on your data load and usage.

**Built-in alerting, machine learning, and correlation features:**  
Splunk includes features for setting alerts, detecting anomalies, and uncovering patterns using ML toolkits.

**Disadvantages:**

1. **Costly Licensing**: Splunk can be expensive, especially for large-scale use or high data volumes.
2. **Complex Learning Curve:** Its advanced features and query language can have a steep learning curve, requiring dedicated training.
3. **Resource Intensive:** Splunk can demand significant computing resources and hardware, impacting performance and scalability.
4. **Limited Free Version:** The free version has data indexing limitations, potentially restricting larger organizations' needs.
5. **Maintenance Overhead**: Managing and maintaining Splunk might require dedicated administration and resources, adding to operational overhead.

Splunk emerges as a comprehensive and powerful solution for real-time data analytics, offering intuitive interfaces and robust features that cater to a wide range of industries and use cases. Its ability to manage both structured and unstructured data makes it invaluable for operational intelligence, security, and performance monitoring. When deployed on AWS, Splunk becomes even more versatile offering a scalable, flexible, and cost-effective environment to integrate with services like CloudTrail, CloudWatch, and VPC logs, making it ideal for both enterprise operations and academic research.

### **Technical Requirements:**

* **AWS Account** (Free Tier eligible)
* **Basic familiarity with AWS Console** (launching EC2, setting security groups)
* **Web browser** (Google Chrome or Firefox recommended)
* **Local system with at least 4 GB RAM** (to access the Splunk UI smoothly)
* **Open port 8000** in the EC2 security group (to access Splunk Web)

### **2. Software and Platform:**

* **Splunk Enterprise AMI** from the AWS Marketplace  
  Link: [Splunk Enterprise on AWS Marketplace](https://aws.amazon.com/marketplace/pp/prodview-lcl67he7x7g3u)
* **Amazon EC2 Instance**  
  Recommended: t2.medium or higher (2 vCPU, 4 GB RAM)

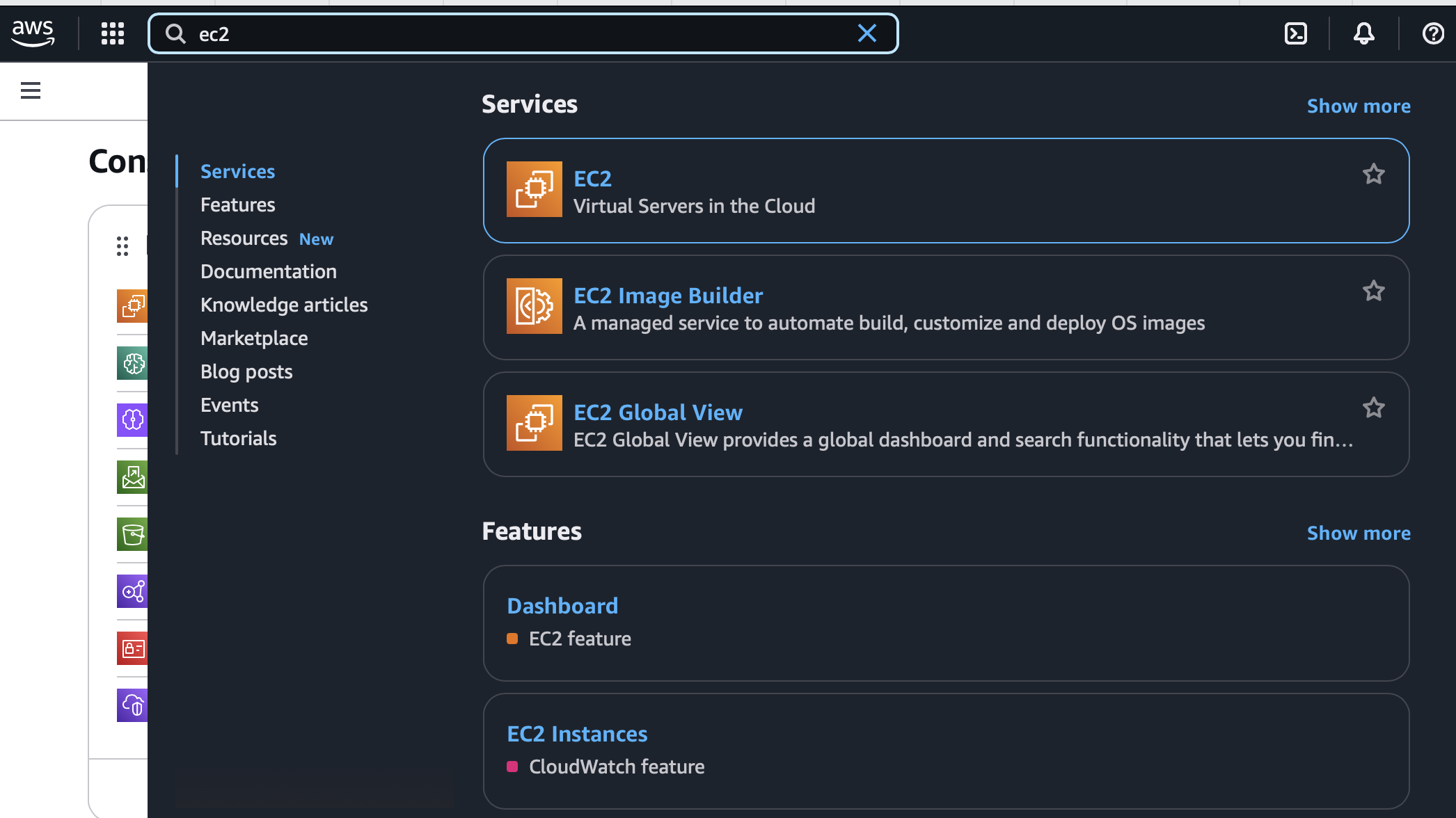
### **3. Estimated Time:**

* ⏱️ **Tutorial Reading Time**: 15–20 minutes
* ⏱️ **Hands-On Setup & Execution**: 45–60 minutes

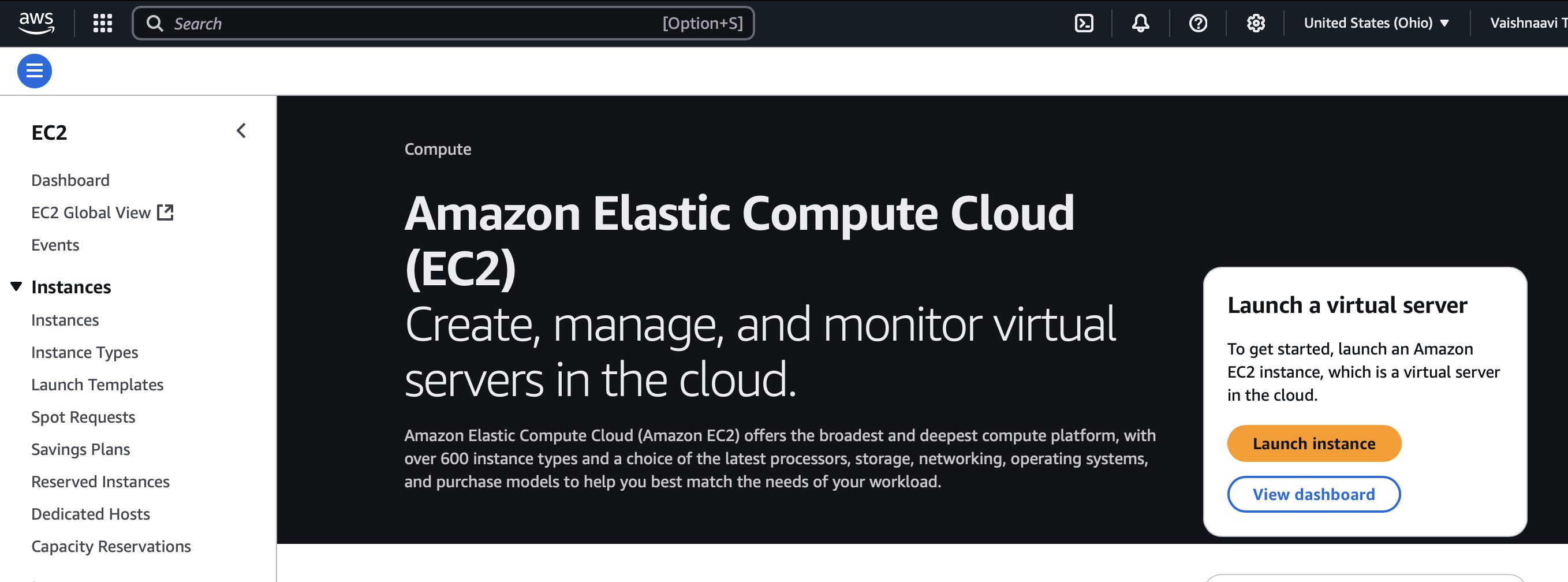
### **4. Estimated Cost:**

* **Splunk Trial**: Free 60-day enterprise trial from AMI
* **AWS EC2 Free Tier**: Covers t2.micro instances (limited performance)
* **Estimated Cost (non-free tier)**: ~$0.05–$0.10/hour for t2.medium  
  (Be sure to stop or terminate your instance after the tutorial to avoid charges)
* **Steps to Follow:**
* **Set-up:**

**1. Login to your AWS account and search for EC2 in services**



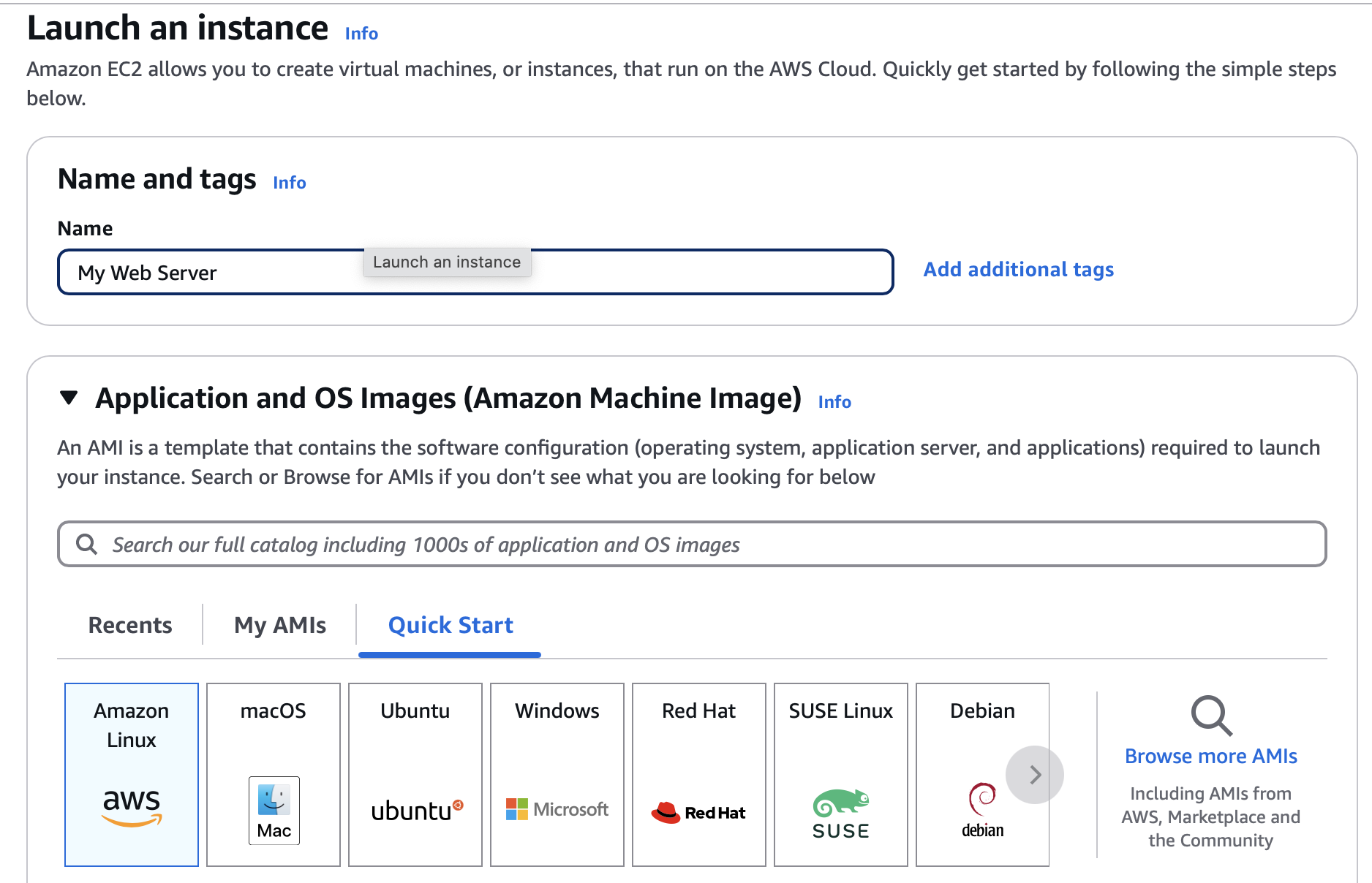
1. **Launch Instance**



**2.1. Give a name to your instance**

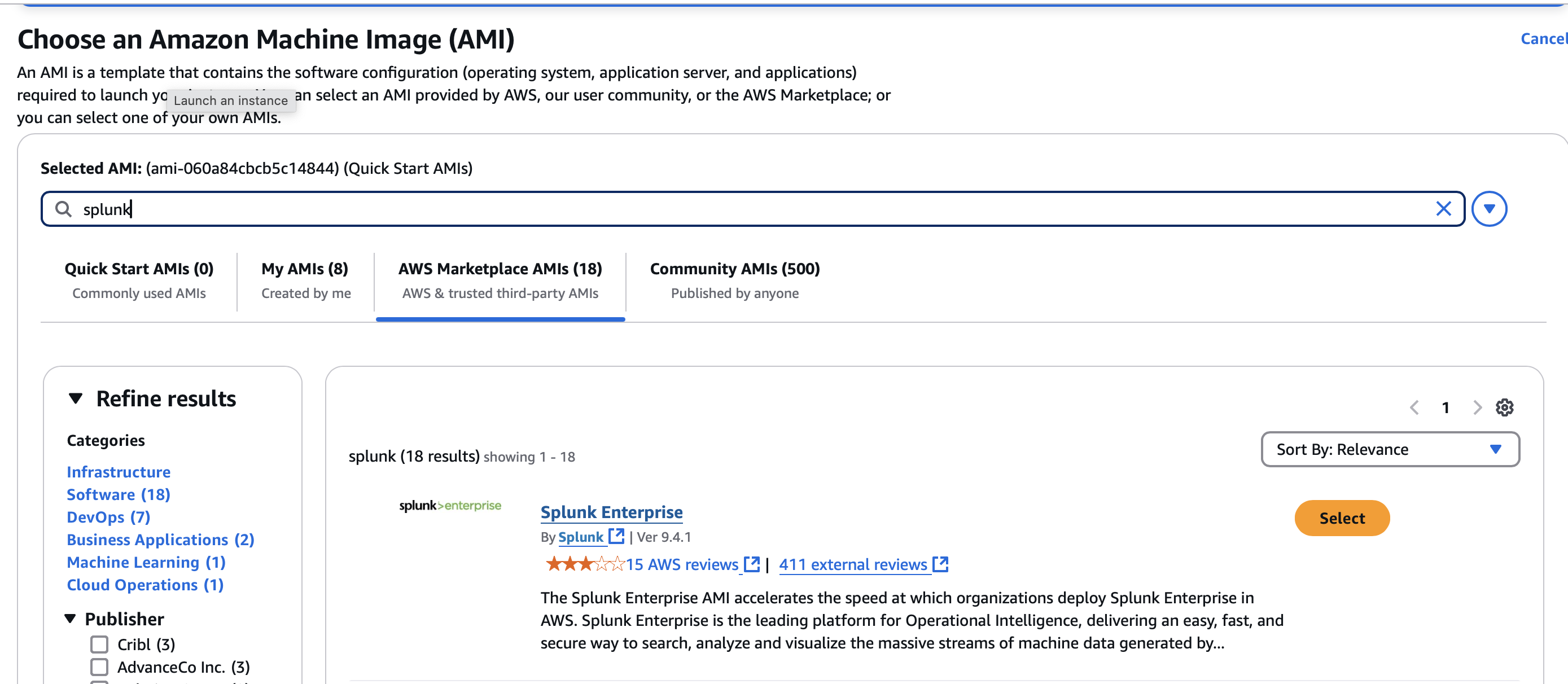
**Application and OS Images**

Click on Browse for more AMIs

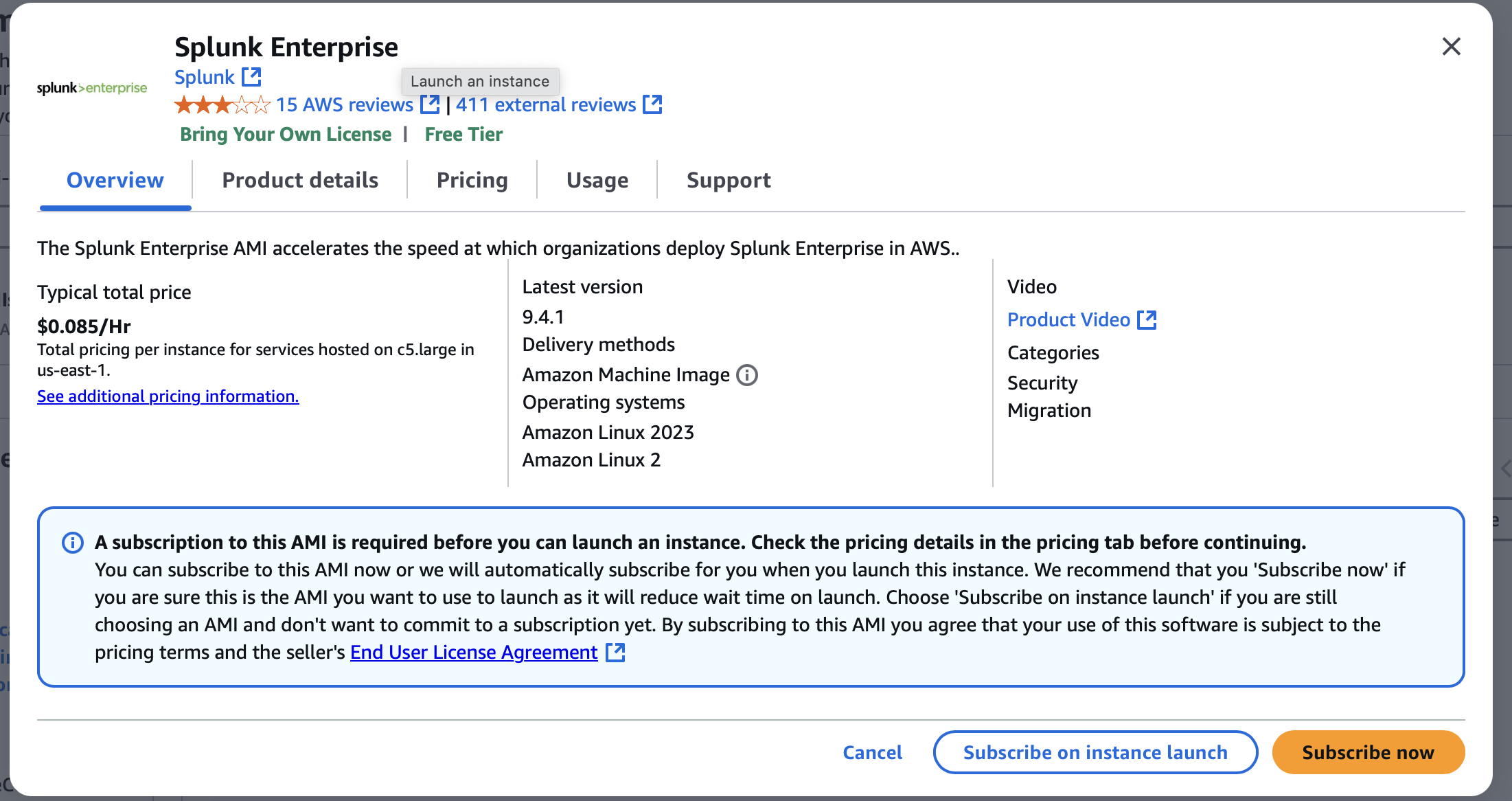


**2.3. Search SPLUNK Enterprise**

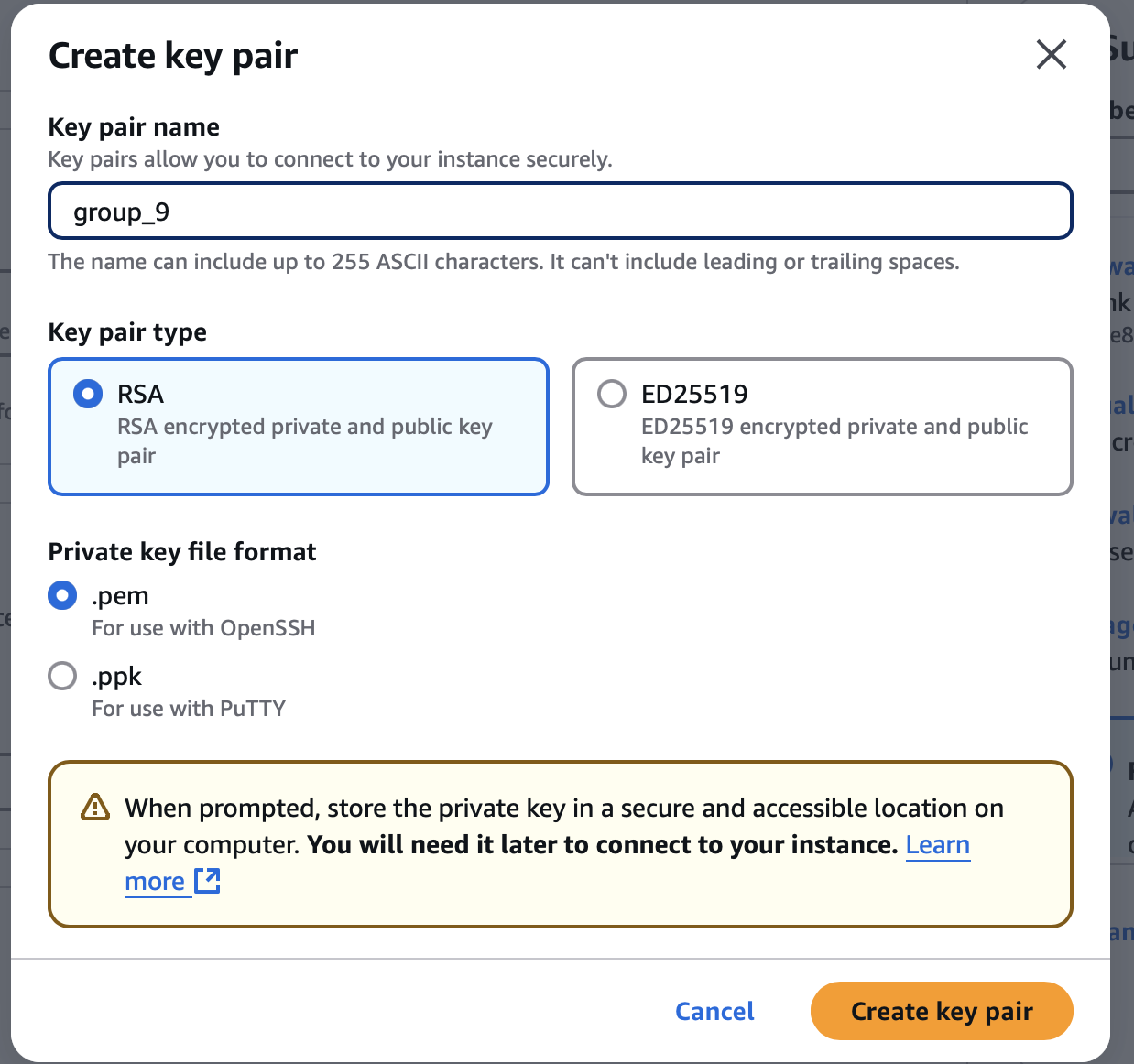
Click on the ‘AWS Marketplace AMIs’ tab and select ‘SPLUNK Enterprise’



**2.4. Subscribe for SPLUNK Enterprise**

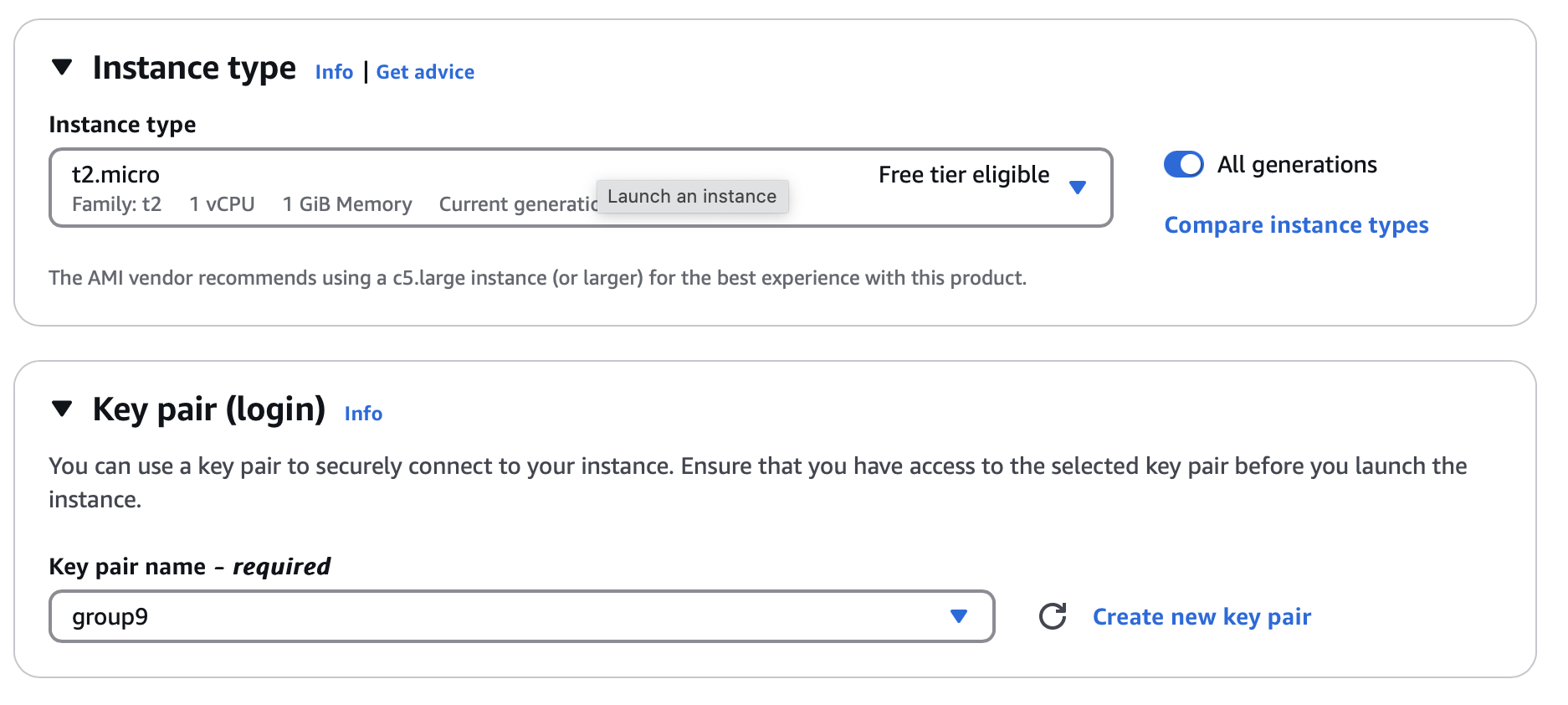


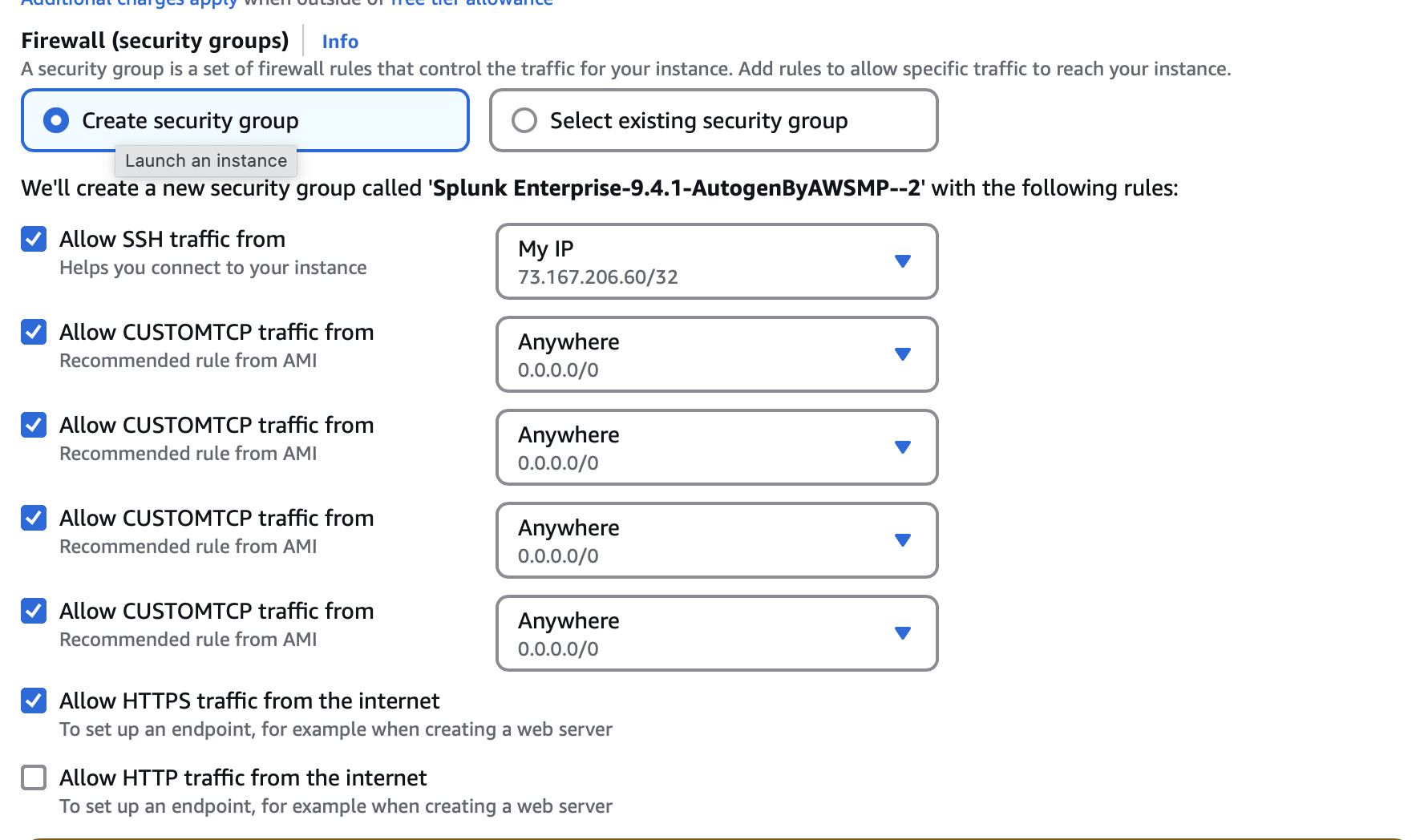
* 1. **Creating a key pair**



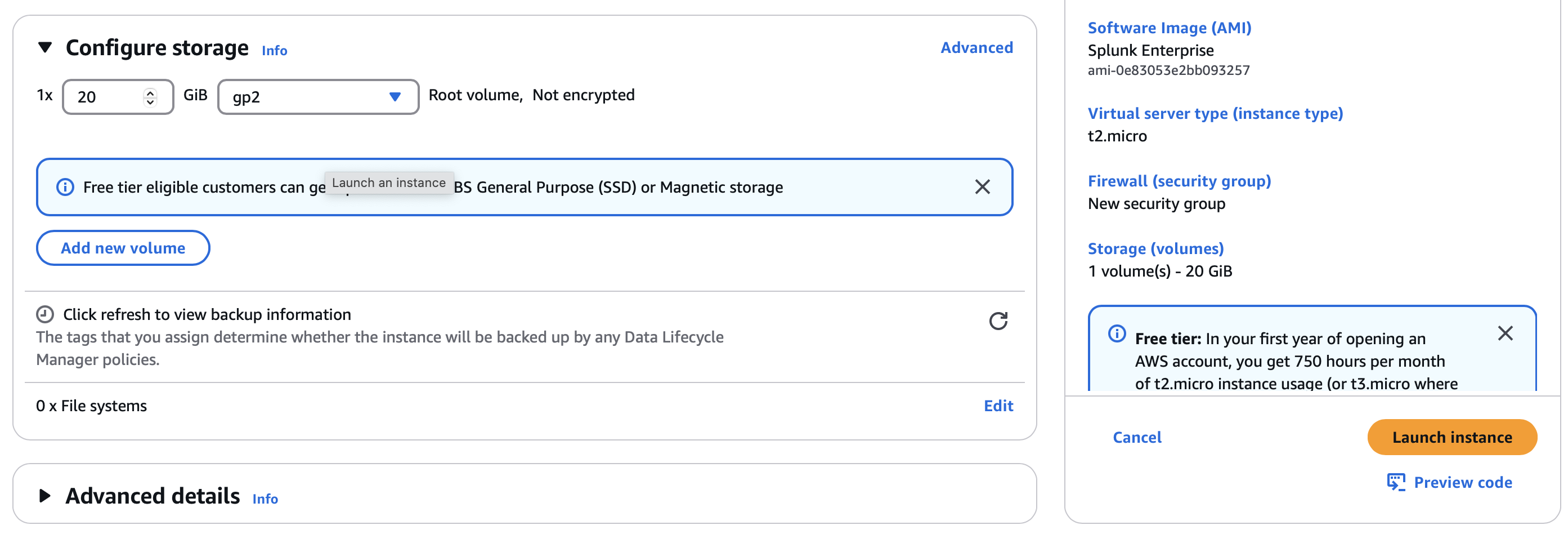
**2.6. Instance type**

Select t2.micro from the drop down

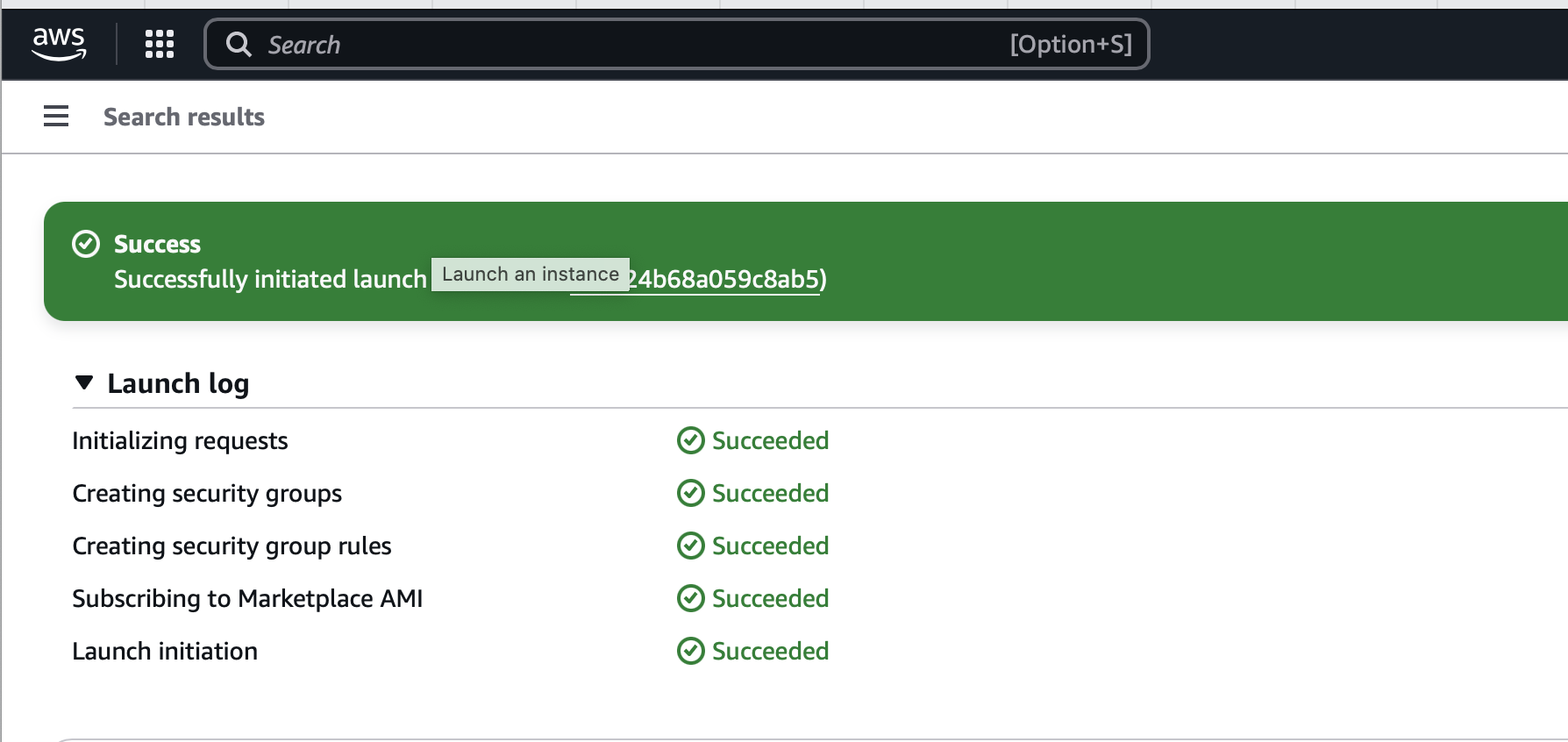


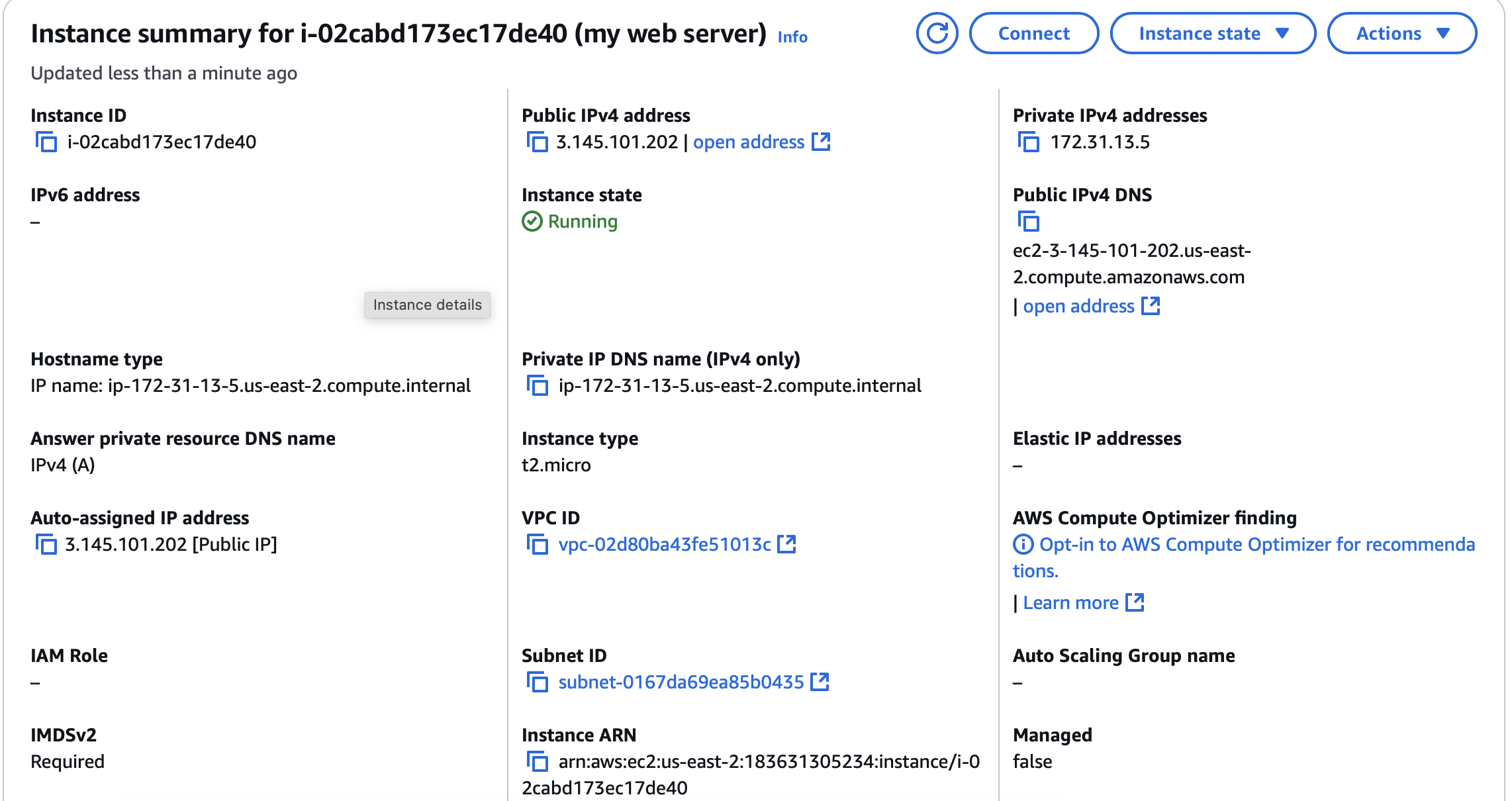


**2.7 Launch the instance**



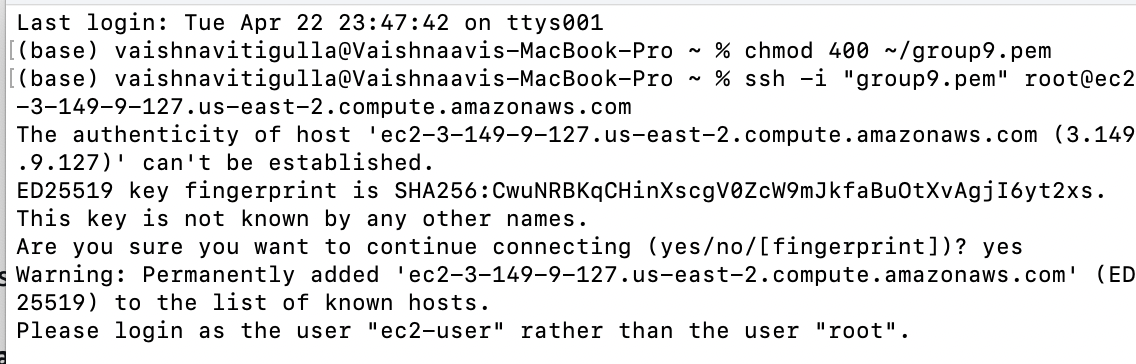
`

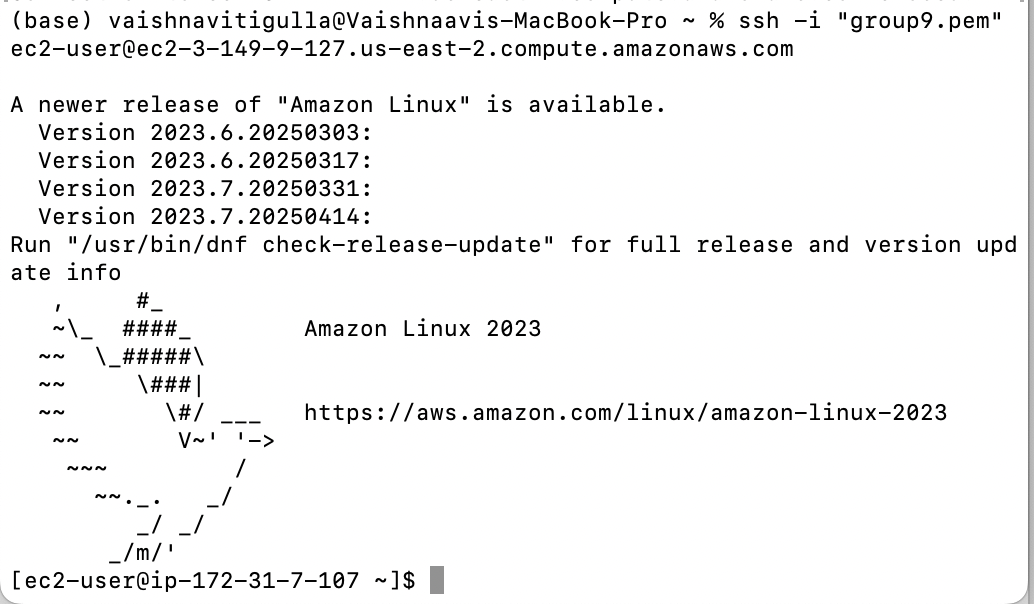




**3. Connection to Ubuntu:**

Open Ubuntu and connect SSH





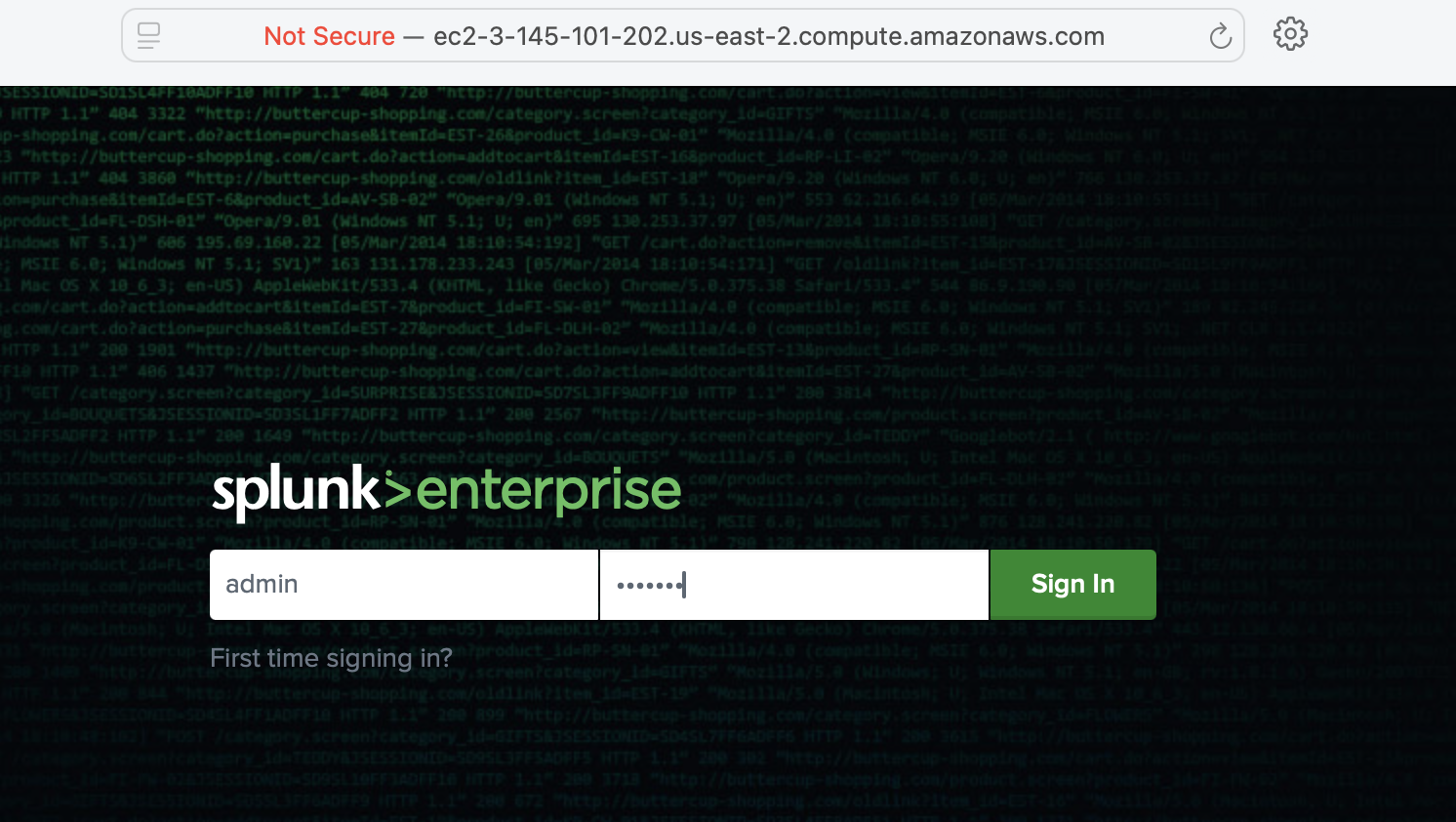


<http://ec2-3-149-9-127.us-east-2.compute.amazonaws.com:8000>

<http://ec2-3-145-101-202.us-east-2.compute.amazonaws.com:8000>

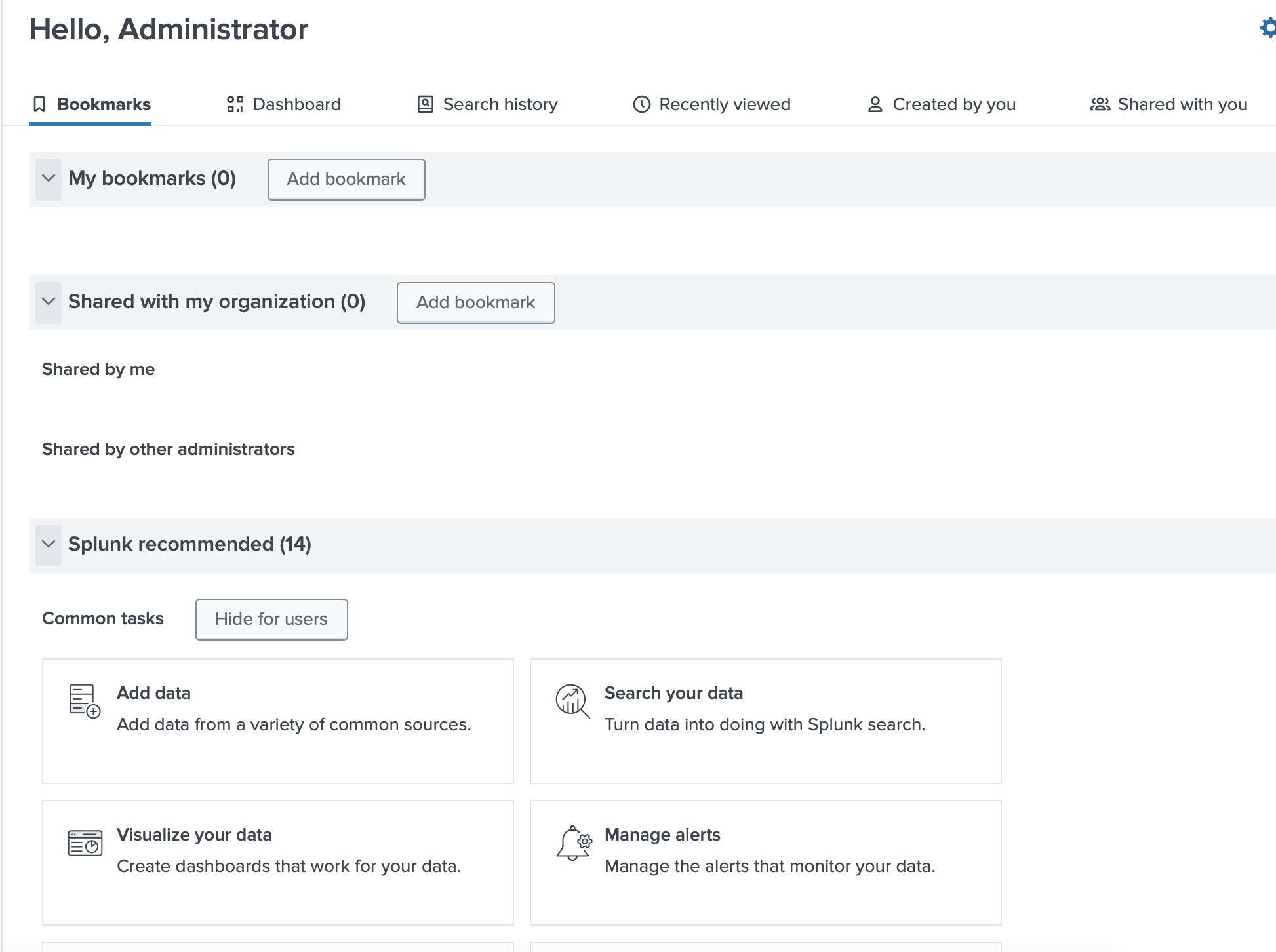
**4. SPLUNK:**

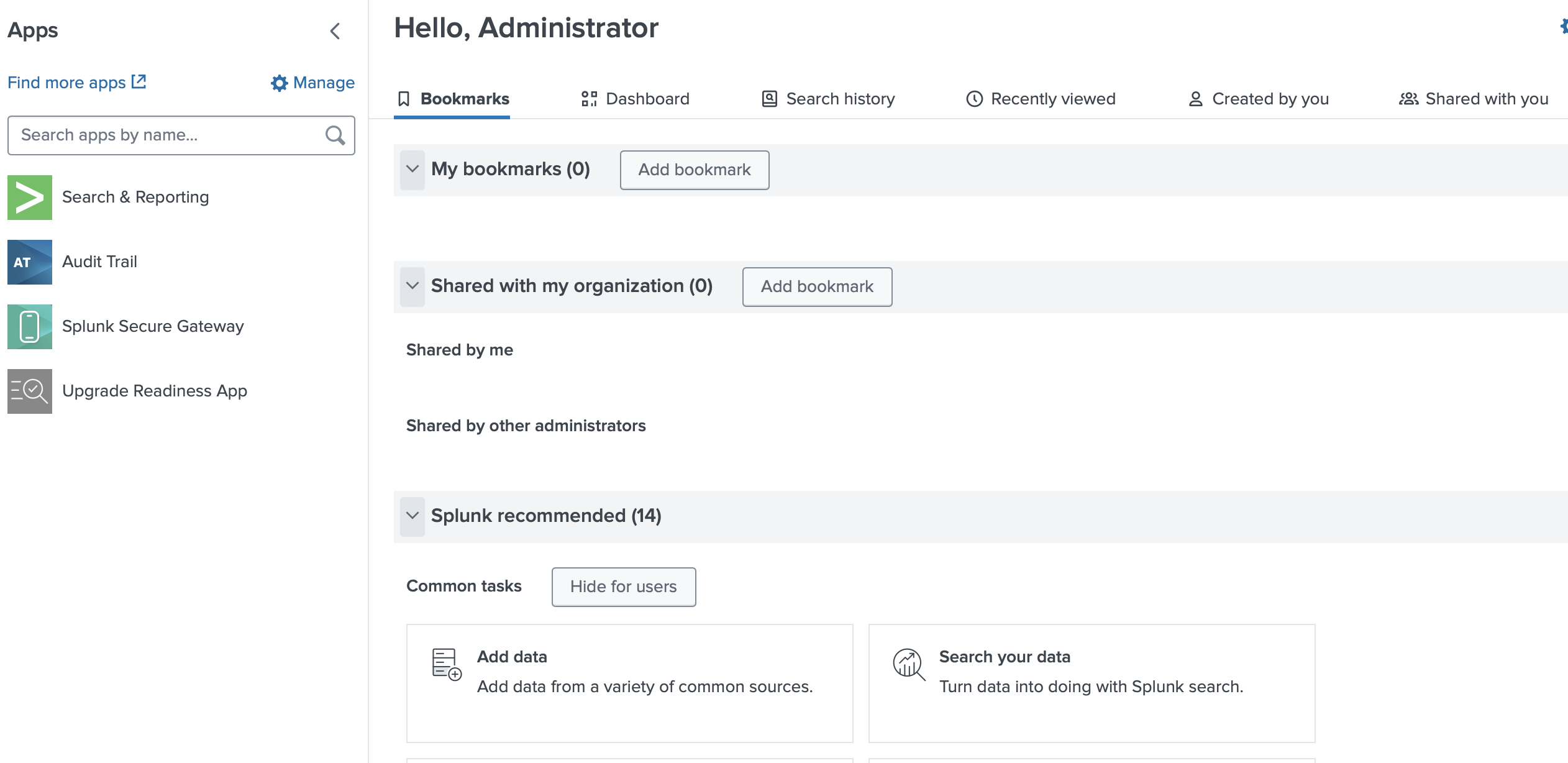
**4.1. Login:**



**4.2. Activity > add data >**

Follow the tutorial for add your dataset (CSV file)

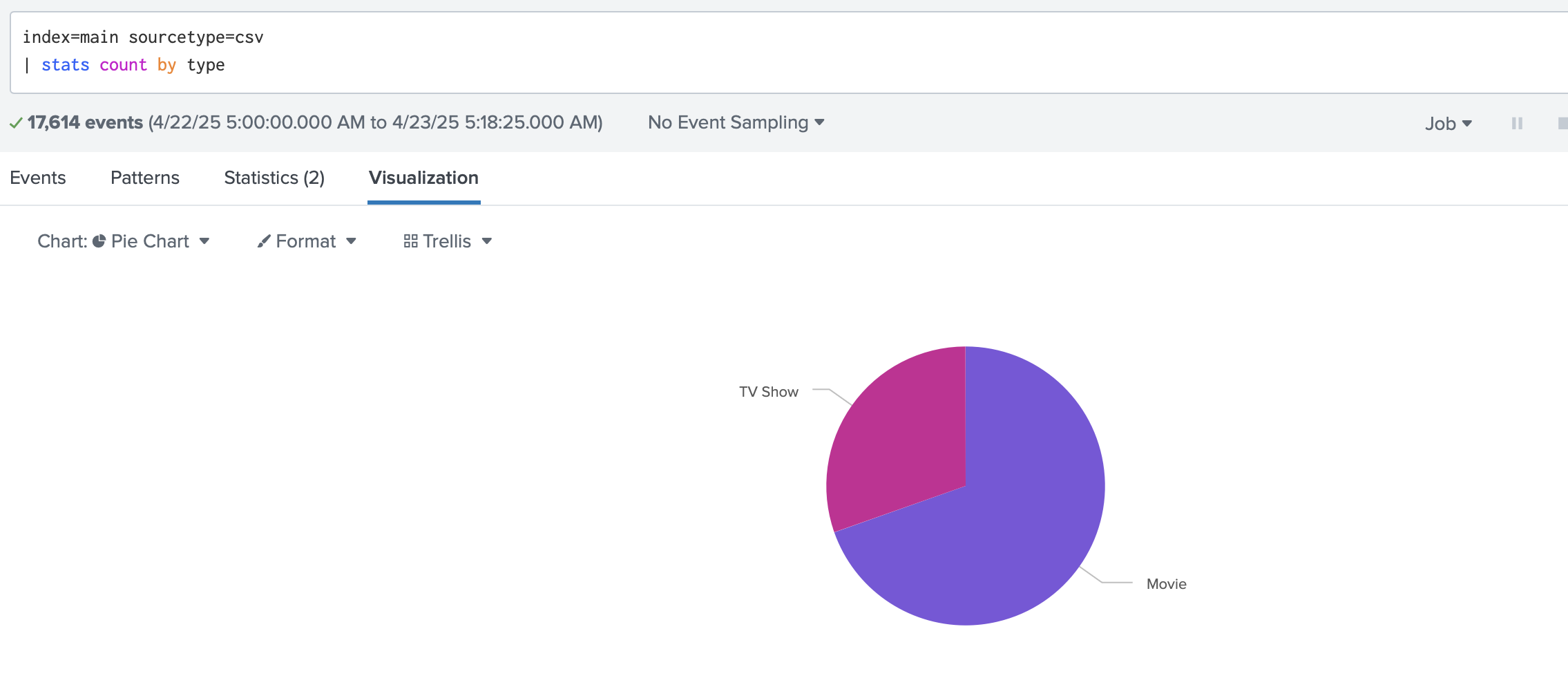


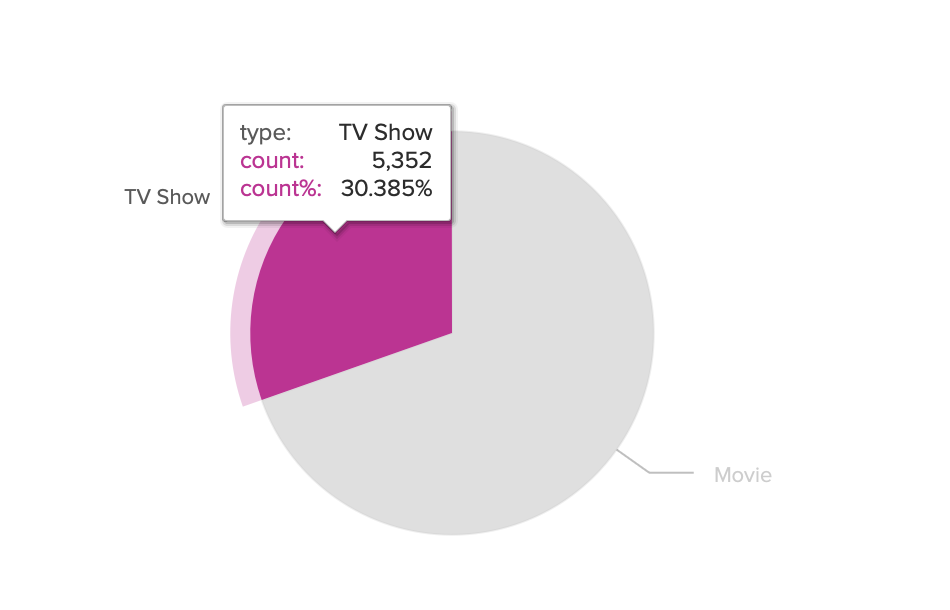


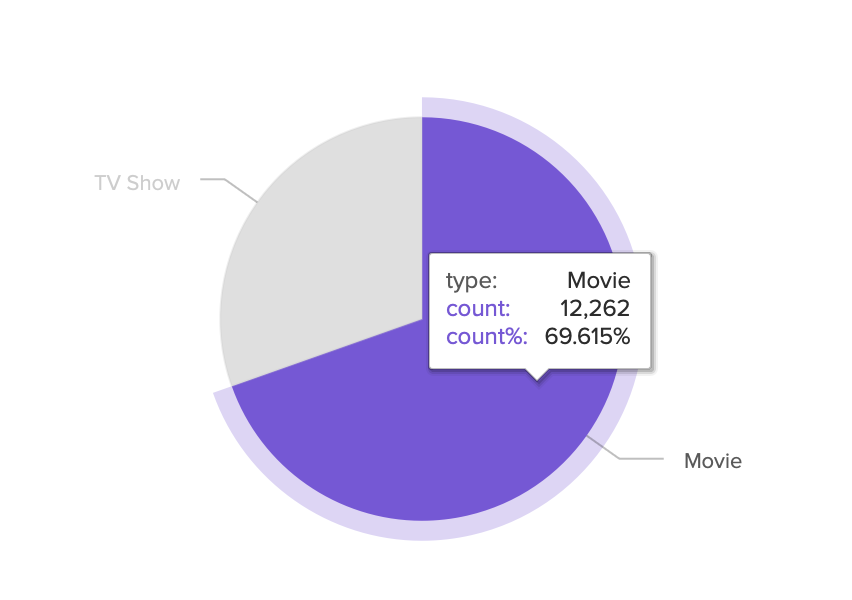
**4.3. Visualization:**

source= “Netflix.csv” , index= Netflix source type=csv

1. **Movies vs TV Shows**

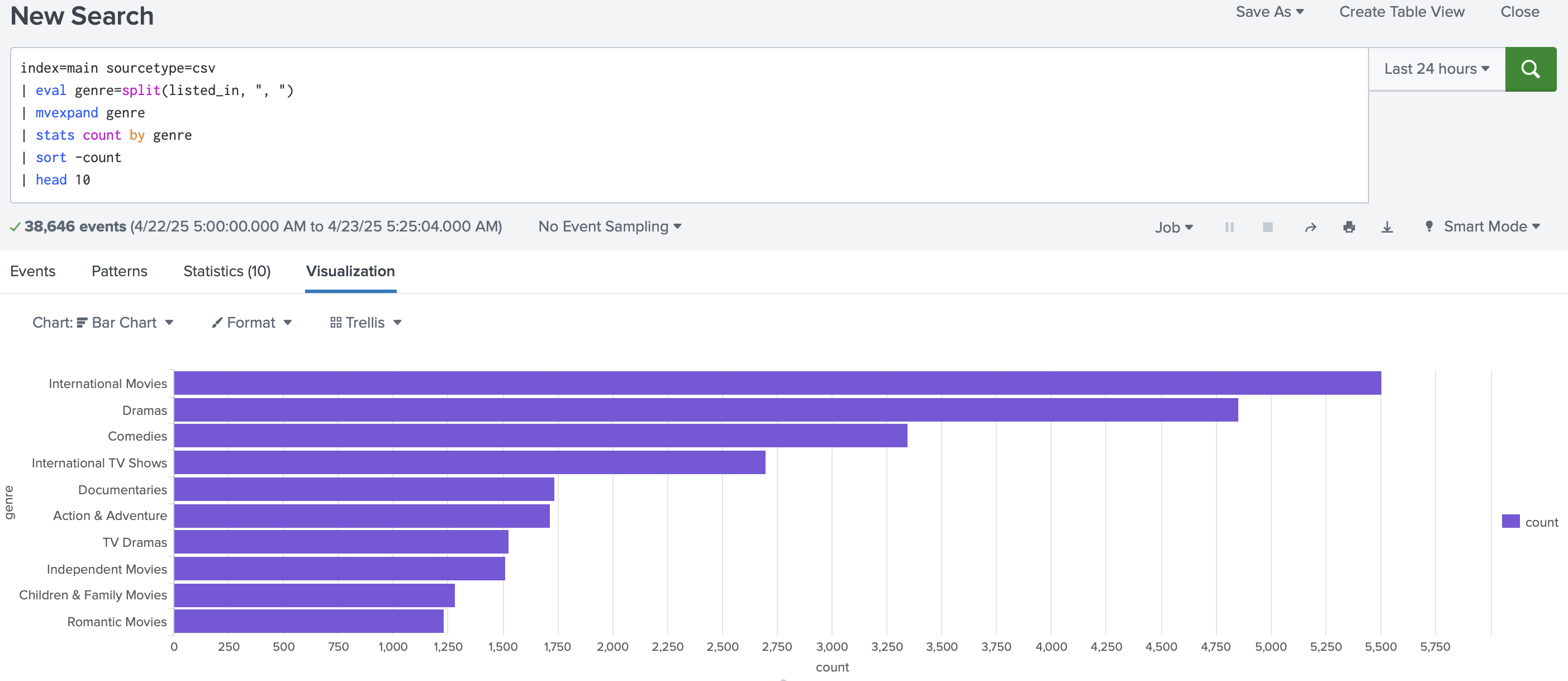
1111





**2.Genres**

source= “Netflix.csv” , index= Netflix source type=csv |



**3 Top 10 Countries by Content**

source= “Netflix.csv” , index= Netflix source type=csv |

