A PRIVACY SEARCH ENGINE

Project Report

01 JANUARY 2022

SIST

PROFESSIONAL TRAINING REPORT

at

Sathyabama Institute of Science and Technology (Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

NAME: Yasar Arafat E
(Reg.No: 40111467)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU

NOVEMBER 2022

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
(Established under Section 3 of UGC Act, 1956)
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Yasar Arafat E (40111467)** who carried out the project entitled "**A privacy search engine**" under my supervision from Aug 2022 to Oct 2022.

Internal Guide

Dr. A. C. Santha Sheela

Head of the Department

Dr.L.Lakshmanan M.E., Ph.D

Submitted for Viva voce Examination held on ______________________

Internal Examiner **External Examiner**

DECLARATION

I Yasar Arafat E hereby declare that the Project Report entitled A privacy Search Engine done by me under the guidance of Dr. A. C. Santha Sheela is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing **Dr.L.Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide
Dr.A.C. Santha Sheela for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS

(i) **MSIX** - Metered Services Information Exchange(Windows app package format)

(ii) **RBI** - Remote Browser Isolation

(iii) **AVD** – Azure Virtual Desktop

(iv) **VM** – Virtual Machine

(v) **Azure AD** – Azure Active Directory

(vi) **ADE** - Azure Disk Encryption

(vii) **AKV** - Azure Key Vault

(viii) **ABS** - Azure Blob Storage

(ix) **ADO** - Azure DevOps

(x) **ACS** - Azure Cloud Switch

Chapter - 1 INTRODUCTION

A Web search app made to do searches in the Internet, to find anything we want with freedom and privacy. This search engine has the mission to make searches truly free, putting people in control of the search algorithms — and not the inverse. It proposes a new Web search experience, being an open and collaborative search tool, to encourage choice, diversity and discoverability on the Web.

It uses ground-breaking pixel streaming technology to execute all web content within a completely safe isolation environment before serving up clean web pages, apps, and content to the user—defending against advanced threats and preventing sensitive data loss.

It offers the following benefits:

- Encrypted Connection(No IP Address Tracking)
- Block Price Trackers
- Avoid Retargeting Ads
- Browse Anonymously
- Receive Unprofiled news
- Prevent Online Profiling
- enables users to access websites without worrying about downloading malicious webpages even if their browsers are outdated, vulnerable or have insecure plugins installed.

Chapter - 2 Aim and Scope of the present investigation

While search engines have become vital tools for searching information on the Internet, privacy issues remain a growing concern due to the technological abilities of search engines to retain user search logs. Although such capabilities might provide enhanced personalized search results, the confidentiality of user intent remains uncertain. Studies have noted that web search query confidentiality continues to be a difficult problem, mainly due to the monetization of search results by search engines.

For instance, on the rationale for retaining user web search query logs, search engine companies offer the following reasons for doing so
(i) Enhancing ranking algorithms,
(ii) Query fine-tuning,
(iii) Improving personalized query results,
(iv) Combating fraud and abuse,
(v) Enabling shared data for research, and
(vi) Enabling shared data for marketing and other commercial purposes.
It is interesting to note that each of the mentioned reasons for retaining user search query logs is a privacy concern.

Even when organizations claim to privatize web search query logs, errors can still be made; as was the case with the 2006 AOL scandal in which a user was re-identified and traced to their geo-location after an anonymized set of web search query logs was published.

In 2009, there was a bug in Google docs that potentially leaked 0.05% of all documents stored in the service. 05% of 1 billion users is 500,000 people.
Another fact worth noting is that Google’s Chrome browser is a potential nightmare when it comes to privacy issues. All user activity within that browser can then be linked to a Google account. If Google controls your browser, your search engine, and has tracking scripts on the sites you visit, they hold the power to track you from multiple angles.

Therefore, the user-based privatization techniques that do not require user data collection are urgently needed.

2.1 IDEATION

A Research study, Cooper (2008) noted that web search query obfuscation techniques could be judged using the following criteria

(i) Effectiveness of the method to protect user privacy,

(ii) Effectiveness of the procedure to conserve the usefulness of query results, and

(ii) How effectively the user can have control to implement the privacy technique

The idea was inspired by the concept of RBI(Remote Browser Isolation)
Browser isolation is a technology that keeps browsing activity secure by separating the process of loading webpages from the user devices displaying the webpages. This way, potentially malicious webpage code does not run on a user’s device, preventing malware infections and other cyber attacks from impacting both user devices and internal networks.
Visiting websites and using web applications involves a web browser loading content and code from remote, untrusted sources (e.g. faraway web servers), then executing that code on a user's device. From a security perspective, this makes browsing the web a fairly dangerous activity. Browser isolation instead loads and executes code far away from users, insulating them and the networks they connect to from the risks — similar to how using robots to perform certain dangerous tasks within a factory can keep the factory workers safer.
Browser isolation can be an important component of a Zero Trust security model, in which no user, application, or website is trusted by default.

2.2 Already Proposed Solution

There are some so-called privacy search engines which claims to make your search queries anonymous but in the end they are all for-profit companies and they all need sources of income to survive. Search engines cost money to maintain. Their increasingly powerful algorithms are the result of many man hours (and processing power) which all cost huge amounts of money. In return for access to vast amounts of information we are asked to tolerate the search engine companies use of our data.

Chapter – 3 Technologies and software Used

Azure is one of the main component used in creating the search engine software. Microsoft Azure, a public cloud computing platform **provides a range of cloud services, including compute, analytics, storage and networking.**

Web technologies like HTML, CSS, JS are used for creating the WebSearch Query Interface.

Customsearch.ai, an Custom Search API is used for the ranking of the websites, global-scale search index, and document processing ability.

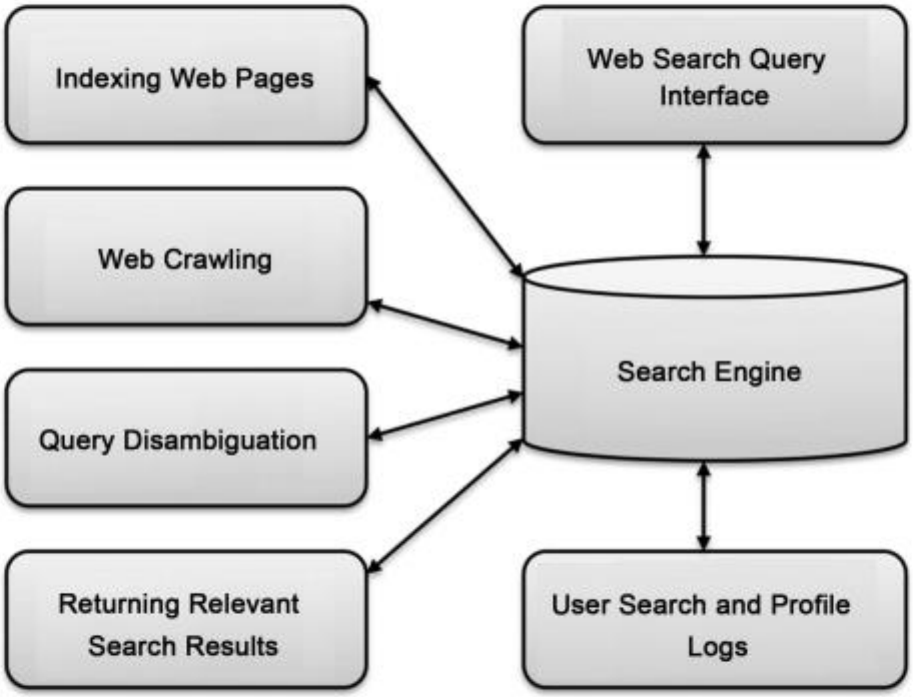

Indexing Web Pages
Web Search Query Interface
Web Crawling
Search Engine
Query Disambiguation
Returning Relevant Search Results
User Search and Profile Logs

Fig. 3.1

Software Layers

- WebSearch Query Interface (Built primarily with web technologies like HTML, CSS, JS etc..)
- Customsearch.ai(Custom Search API gives the powerful ranking, a global-scale search index, and document processing ability.)
- Azure Cloud services

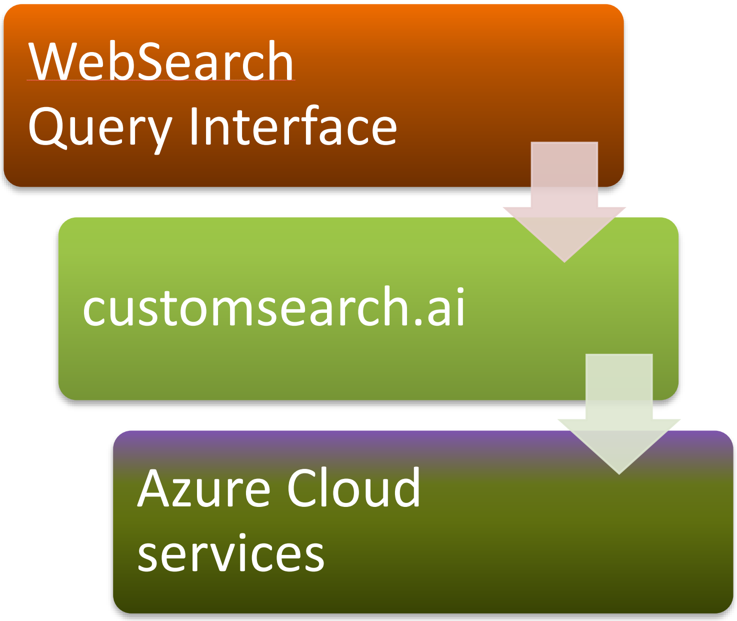


Fig. 3.2

3.1 CustomSearch.ai

The core technology of the software works in
four steps — by identifying on-topic sites and images,
providing automatic query suggestions,
applying the Bing ranker, and
delivering relevant search results.
The parameters can be adjusted and usage insights can be accessed anytime.

The parameters of the search domain can be defined and adjusted Quickly and reliably. Search result layouts can be rendered using a hosted Bing user experience.

Automatic query suggestions can be predefined for the content within selected domains. It can help users complete their queries faster by adding intelligent type-ahead capabilities to the application.

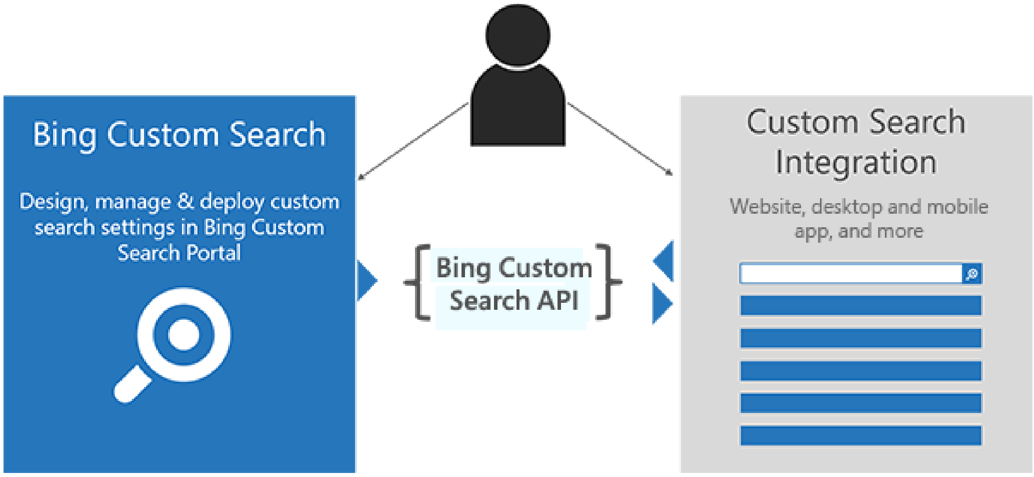

Bing Custom Search
Design, manage & deploy custom search settings in Bing Custom Search Portal
Bing Custom Search API
Custom Search Integration
Website, desktop and mobile app, and more

Fig. 3.3

Tailored search for specific content

A compelling search experience for content, such as web pages, images, videos, and even autosuggested words. With Bing Custom Search, client can promote or filter results to drive the outcome they want.

Complete control over search results

Control ranking, pinning, blocking, boosting, and demoting sites.

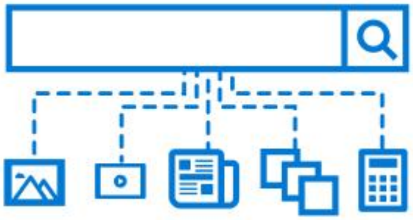


Fig. 3.4

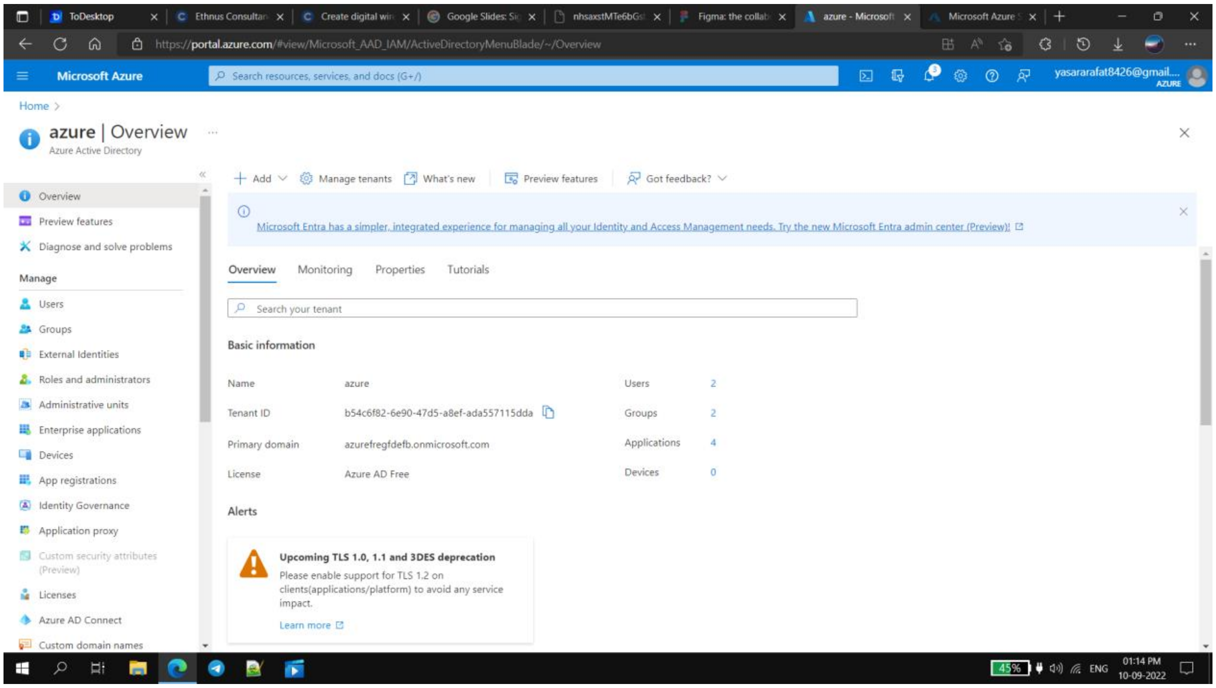
3.2 Azure

Azure cloud computing service is used in the Search Engine software. It is responsible for managing the backend, server load and application processing.

A resource group container is created in the Azure cloud to manage the global-scale search index, document processing and client side web search query result processing.

A bing resource is created under the Resource group and it is used to completely tailor and maintain the search engine. The web search query insights can be seen in the resource group and it can be used to improve the application.

STEP 1: An Azure Active Directory is created



azure | Overview

Azure Active Directory

Microsoft Entra has a simpler, integrated experience for managing all your Identity and Access Management needs. Try the new Microsoft Entra admin center (Preview)!

Basic information

Name	azure	Users	2
Tenant ID	b54c6f82-6e90-47d5-a8ef-ada557115dda	Groups	2
Primary domain	azurefregfdefb.onmicrosoft.com	Applications	4
License	Azure AD Free	Devices	0

Alerts

Upcoming TLS 1.0, 1.1 and 3DES deprecation

Please enable support for TLS 1.2 on clients(applications/platform) to avoid any service impact.

Learn more

Fig. 3.5

Step 2: An Azure Virtual Desktop is created under the Azure Active Directory.

Virtual Desktops are VDI-based desktops with a unique set of configuration and Operating Systems. These desktops are virtual machines with dedicated resources, applications, and user settings.

Step 3: A host pool is created inside the Azure Virtual Desktop and a session host is created in the host pool.

Step 4: A virtual network and subnets is created for a private and secure network for the session host.

Step 5: A network security group and network interface is created with the Inbound port rules to connect with the session host

Step 6: Azure AD Domain Services is used to create an domain with Azure for hosting the virtual machine.

Step 7: An Azure storage disk is created with enough storage and capacity inside the session host to run the applications.

Step 8: An MSIX package of the application is uploaded to the session host.

A desktop application is created for the search engine and an MSIX package is created.

Step 9: The application is connected either using the RDP host or in the web interface.

Step 10: The application is accessed using the RDP host or through the web interface and the queries are searched to crawl websites across the web.

Chapter – 4, Results and Performance Analysis

The msix package was uploaded to the virtual machine and the search engine app was successfully simulated in the web using the remote web servers. The concept is to deliver virtualized app on end-point devices from a remote server setup.

Developing a streaming application needs consideration of several factors like features, tech stack required, and cost analysis. Considering each aspect and providing a higher user experience to your customers need proper planning, execution, testing of the product, monitoring them, maintaining performance, etc.

The performance of the streamed application is great in terms of security prospects. However, there are speed, stability and accessibility concerns surrounding the search engine app becoming main stream for every day uses.

Benefits:

- Dangerous downloads are deleted
- Malicious scripts do not execute on a device or inside a private network
- Zero-day exploits through the browser are blocked
- Malicious web content can be blocked without having to block entire websites

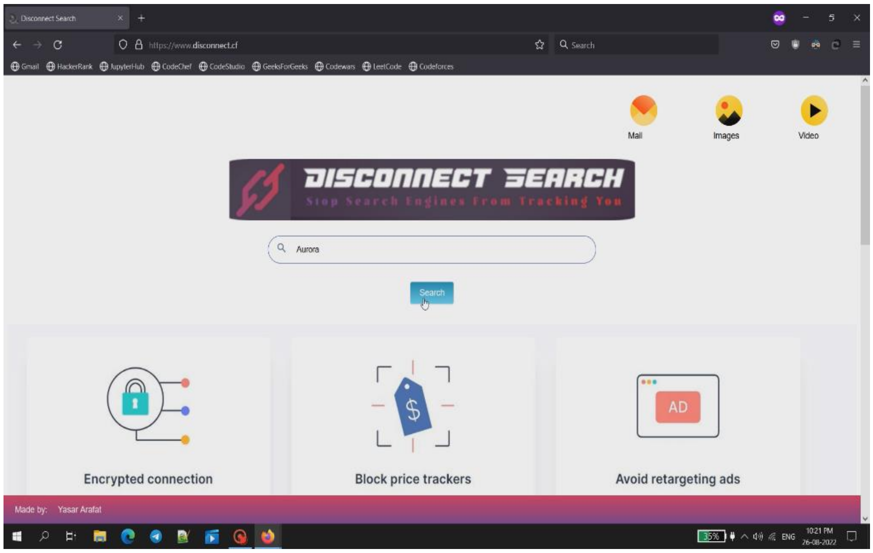
Pros

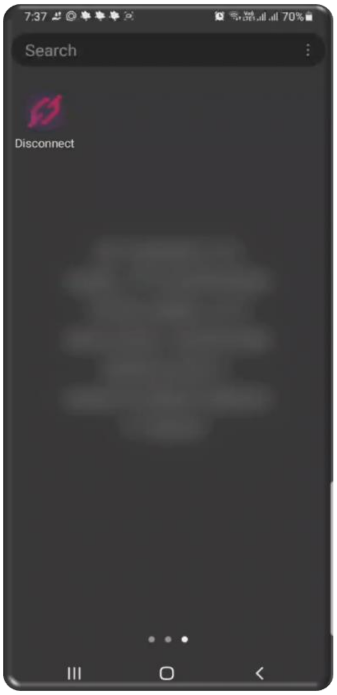
- Anonymous search results
- Not backed up by monopolistic corporate companies
- No Trackers
- No personalized ads

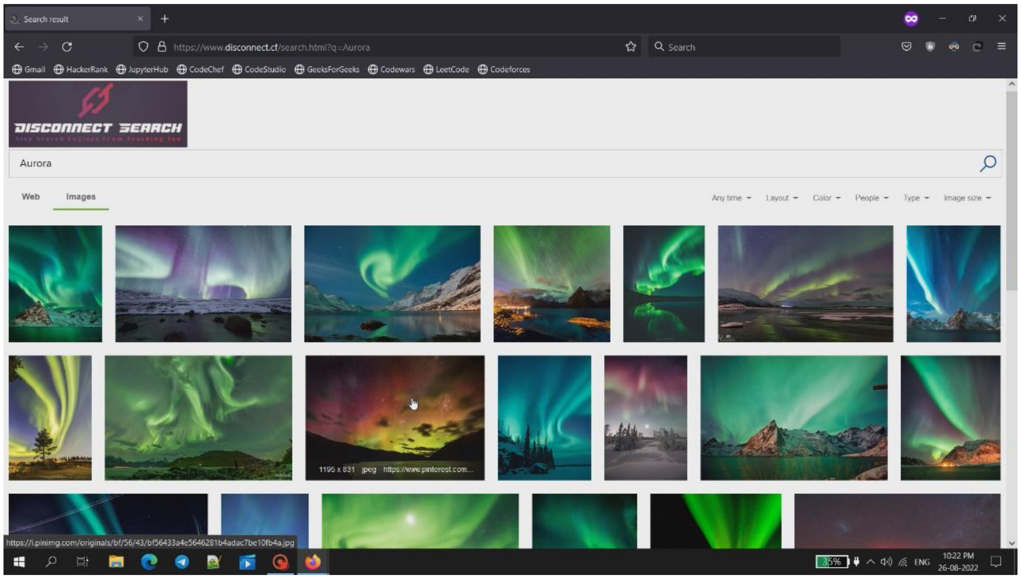
Cons

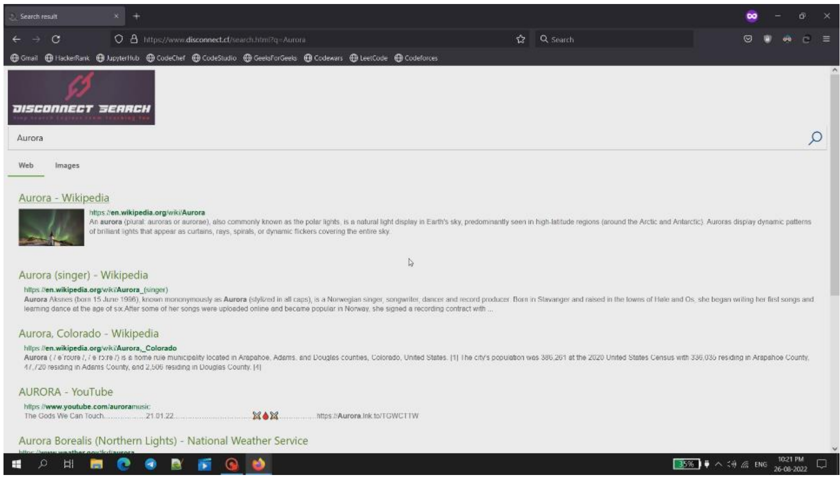
- As it’s focused on privacy and protection of the user data, making a business model is hard
- Less Scalability

4.1 Web and Mobile Application






Search result
https://www.disconnect.cf/search.html?q=Aurora
Search
Gmail HackerRank JupyterHub CodeChef CodeStudio GeeksForGeeks Codewars LeetCode Codeforces
DISCONNECT SEARCH
Aurora
Web
Images
Any time
Layout
Color
People
Type
Image size
1195 x 831 jpeg https://www.pinterest.com...
https://i.pinimg.com/originals/bf/56/43/bf56433a4e5646281b4adac7bc10fb4a.jpg
35%
ENG
10:22 PM
26-08-2022



Aurora

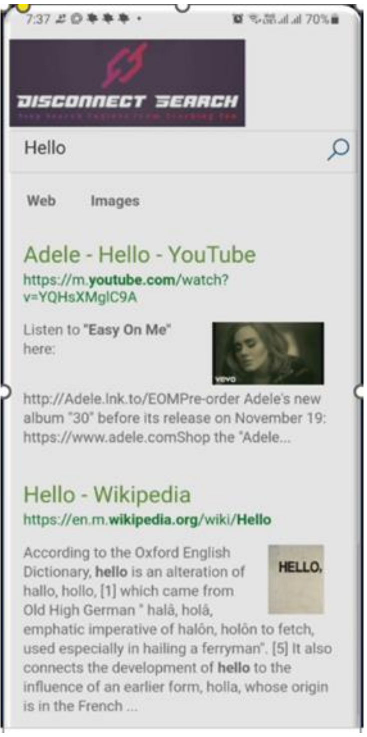
Aurora - Wikipedia
https://en.wikipedia.org/wiki/Aurora
An **aurora** (plural: auroras or aurorae), also commonly known as the polar lights, is a natural light display in Earth's sky, predominantly seen in high-latitude regions (around the Arctic and Antarctic). Auroras display dynamic patterns of brilliant lights that appear as curtains, rays, spirals, or dynamic flickers covering the entire sky.

Aurora (singer) - Wikipedia
https://en.wikipedia.org/wiki/Aurora_(singer)
Aurora Aksnes (born 15 June 1996), known mononymously as **Aurora** (stylized in all caps), is a Norwegian singer, songwriter, dancer and record producer. Born in Stavanger and raised in the towns of Hale and Os, she began writing her first songs and learning dance at the age of six.After some of her songs were uploaded online and became popular in Norway, she signed a recording contract with ...

Aurora, Colorado - Wikipedia
https://en.wikipedia.org/wiki/Aurora,_Colorado
Aurora (/ ə'roʊrə /, / ə'rɔːrə /) is a home rule municipality located in Arapahoe, Adams, and Douglas counties, Colorado, United States. [1] The city's population was 386,261 at the 2020 United States Census with 336,035 residing in Arapahoe County, 47,720 residing in Adams County, and 2,506 residing in Douglas County. [4]

AURORA - YouTube
https://www.youtube.com/auroramusic
The Gods We Can Touch................21.01.22..............................🦋🩸🦋..............https://**Aurora**.lnk.to/TGWCTTW

Aurora Borealis (Northern Lights) - National Weather Service



Hello

Adele - Hello - YouTube
https://m.youtube.com/watch?v=YQHsXMglC9A
Listen to **"Easy On Me"** here:
http://Adele.lnk.to/EOMPre-order Adele's new album "30" before its release on November 19: https://www.adele.comShop the "Adele...

Hello - Wikipedia
https://en.m.wikipedia.org/wiki/Hello
According to the Oxford English Dictionary, **hello** is an alteration of hallo, hollo, [1] which came from Old High German " halâ, holâ, emphatic imperative of halôn, holôn to fetch, used especially in hailing a ferryman". [5] It also connects the development of **hello** to the influence of an earlier form, holla, whose origin is in the French ...

Chapter – 5
Summary and Conclusions

Developing a streaming application needs consideration of several factors like features, tech stack required, and cost analysis. Considering each aspect and providing a higher user experience to your customers need proper planning, execution, testing of the product, monitoring them, maintaining performance, etc.

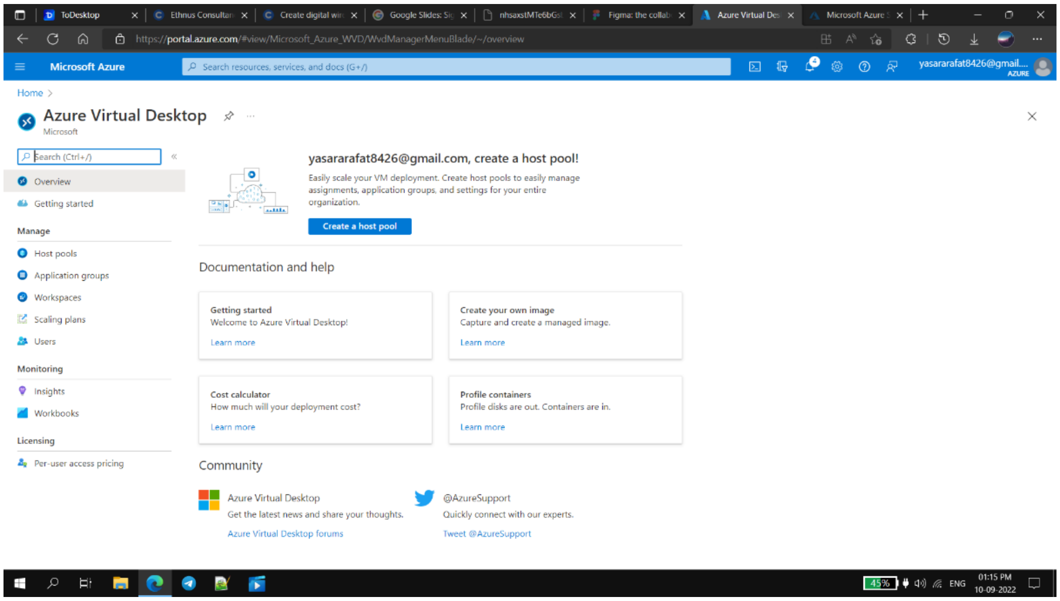
The performance of the streamed application is great in terms of security prospects. However, there are speed, stability and accessibility concerns surrounding the search engine app becoming main stream for every day uses.

References: Scientific journal, scirp.org, Journal of Information Security > Vol.8 No.1, January 2017 by Kato Mivule

Web Search Query Privacy, an End-User Perspective (scirp.org)

APPENDIX

A. SCREENSHOTS



yasararafat8426@gmail.com, create a host pool!

Easily scale your VM deployment. Create host pools to easily manage assignments, application groups, and settings for your entire organization.

Create a host pool

Documentation and help

Getting started
Welcome to Azure Virtual Desktop!
Learn more

Create your own image
Capture and create a managed image.
Learn more

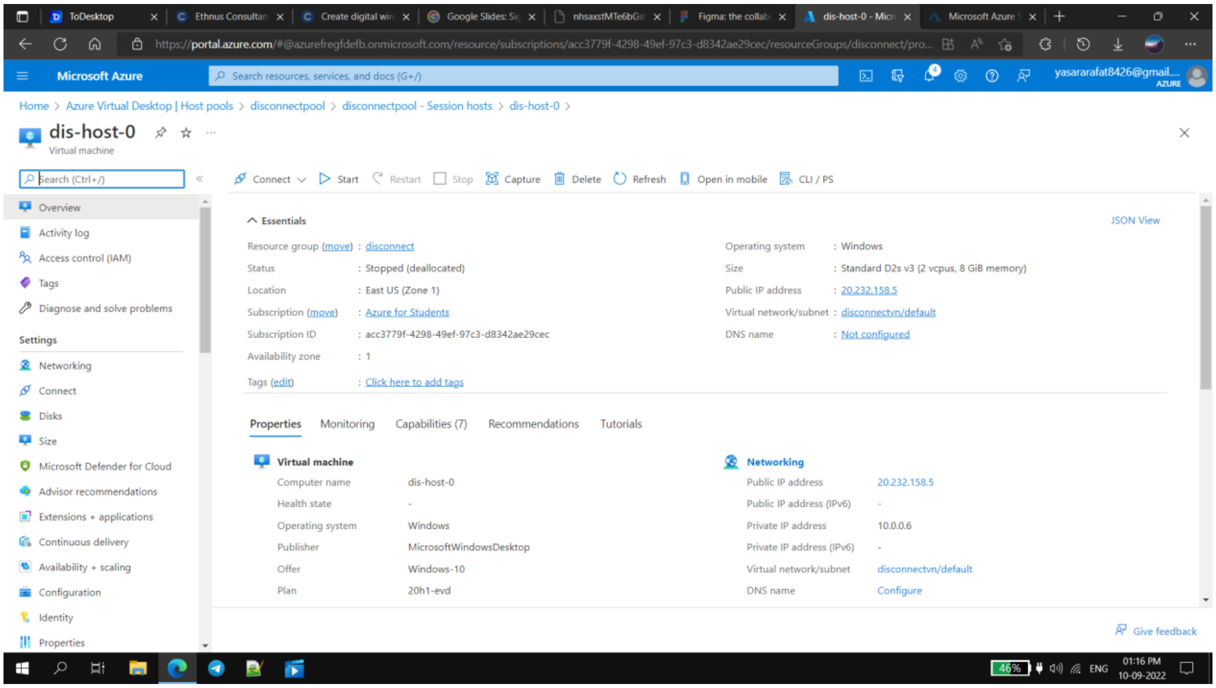
Cost calculator
How much will your deployment cost?
Learn more

Profile containers
Profile disks are out. Containers are in.
Learn more

Community

Azure Virtual Desktop
Get the latest news and share your thoughts.
Azure Virtual Desktop forums

@AzureSupport
Quickly connect with our experts.
Tweet @AzureSupport



https://portal.azure.com/#@azurefregfdefb.onmicrosoft.com/resource/subscriptions/acc3779f-4298-49ef-97c3-d8342ae29cec/resourceGroups/disconnect/pro...

dis-host-0

Virtual machine

Essentials

Resource group (move) : disconnect
Status : Stopped (deallocated)
Location : East US (Zone 1)
Subscription (move) : Azure for Students
Subscription ID : acc3779f-4298-49ef-97c3-d8342ae29cec
Availability zone : 1
Tags (edit) : Click here to add tags

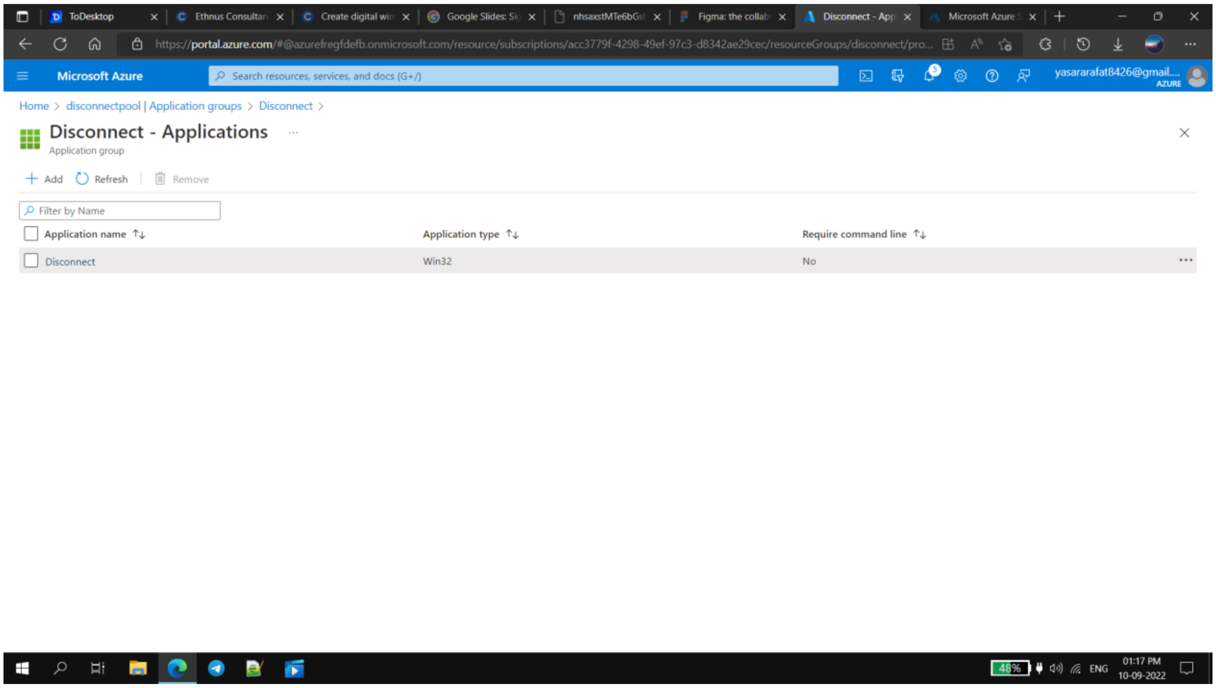
Operating system : Windows
Size : Standard D2s v3 (2 vcpus, 8 GiB memory)
Public IP address : 20.232.158.5
Virtual network/subnet : disconnectvn/default
DNS name : Not configured

Virtual machine

Computer name	dis-host-0
Health state	-
Operating system	Windows
Publisher	MicrosoftWindowsDesktop
Offer	Windows-10
Plan	20h1-evd

Networking

Public IP address	20.232.158.5
Public IP address (IPv6)	-
Private IP address	10.0.0.6
Private IP address (IPv6)	-
Virtual network/subnet	disconnectvn/default
DNS name	Configure

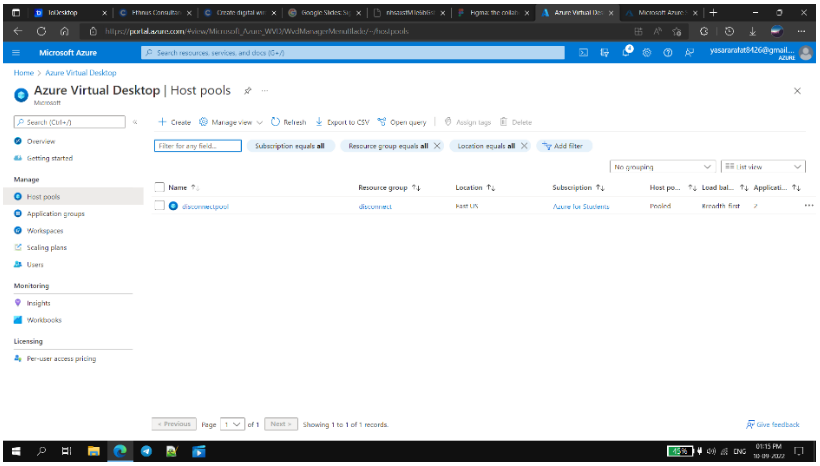

ToDesktop
Ethnus Consultan
Create digital win
Google Slides: Sig
nhsaxstMTe6bGsl
Figma: the collab
Disconnect - App
Microsoft Azure
https://portal.azure.com/#@azurefregfdefb.onmicrosoft.com/resource/subscriptions/acc3779f-4298-49ef-97c3-d8342ae29cec/resourceGroups/disconnect/pro...
Microsoft Azure
Search resources, services, and docs (G+/)
yasararafat8426@gmail....
AZURE
+ Add
Refresh
Remove
Filter by Name
48%
ENG
01:17 PM
10-09-2022

Home > disconnectpool | Application groups > Disconnect >

Disconnect - Applications

Application group

□ Application name ↑↓	Application type ↑↓	Require command line ↑↓
□ Disconnect	Win32	No



Home > Azure Virtual Desktop

Azure Virtual Desktop | Host pools

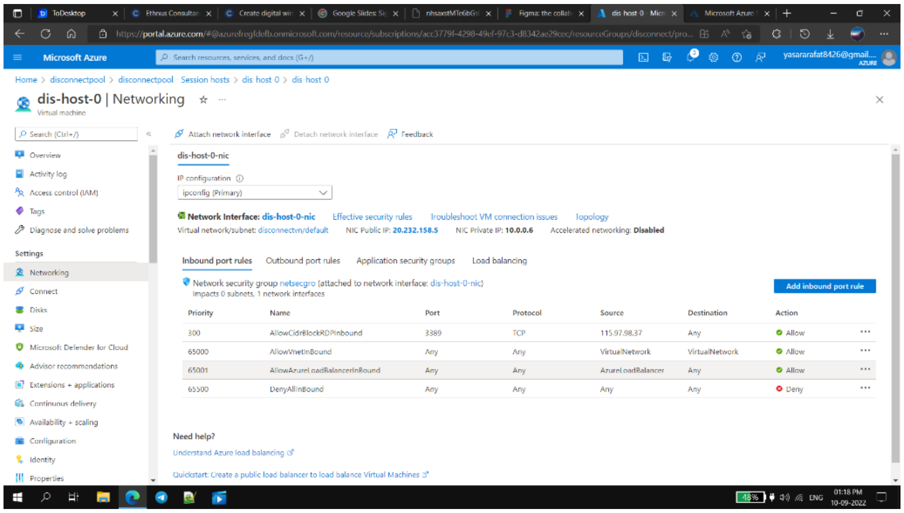
Microsoft

Subscription equals **all** | Resource group equals **all** | Location equals **all** | Add filter

No grouping | List view

Name	Resource group	Location	Subscription	Host po...	Load bal...	Applicati...
disconnectpool	disconnect	East US	Azure for Students	Pooled	Breadth first	2

< Previous Page 1 of 1 Next > Showing 1 to 1 of 1 records.



Home > disconnectpool > disconnectpool Session hosts > dis host 0 > dis host 0

dis-host-0 | Networking

Virtual machine

dis-host-0-nic

IP configuration: ipconfig (Primary)

Network Interface: dis-host-0-nic Effective security rules Troubleshoot VM connection issues Topology

Virtual network/subnet: disconnectvn/default NIC Public IP: **20.232.158.5** NIC Private IP: **10.0.0.6** Accelerated networking: **Disabled**

Inbound port rules | Outbound port rules | Application security groups | Load balancing

Network security group netsecgro (attached to network interface: dis-host-0-nic)
Impacts 0 subnets, 1 network interfaces

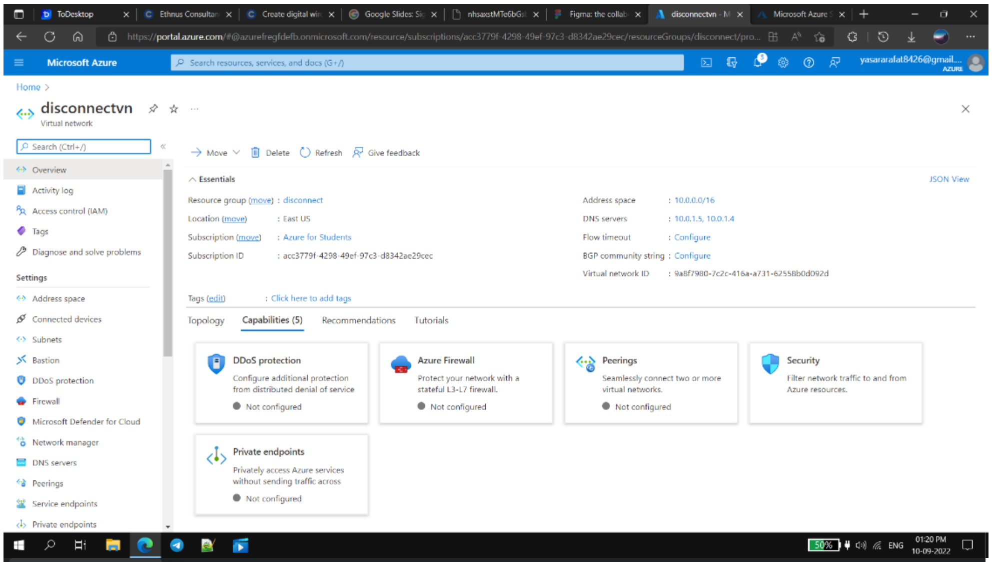
Add inbound port rule

Priority	Name	Port	Protocol	Source	Destination	Action
300	AllowCidrBlockRDPInbound	3389	TCP	115.97.98.37	Any	Allow
65000	AllowVnetInBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow
65001	AllowAzureLoadBalancerInBound	Any	Any	AzureLoadBalancer	Any	Allow
65500	DenyAllInBound	Any	Any	Any	Any	Deny

Need help?

Understand Azure load balancing

Quickstart: Create a public load balancer to load balance Virtual Machines



disconnectvn

Virtual network

Essentials

Resource group (move) : disconnect

Location (move) : East US

Subscription (move) : Azure for Students

Subscription ID : acc3779f-4298-49ef-97c3-d8342ae29cec

Address space : 10.0.0.0/16

DNS servers : 10.0.1.5, 10.0.1.4

Flow timeout : Configure

BGP community string : Configure

Virtual network ID : 9a8f7980-7c2c-416a-a731-62558b0d092d

Tags (edit) : Click here to add tags

DDoS protection

Configure additional protection from distributed denial of service

● Not configured

Azure Firewall

Protect your network with a stateful L3-L7 firewall.

● Not configured

Peerings

Seamlessly connect two or more virtual networks.

● Not configured

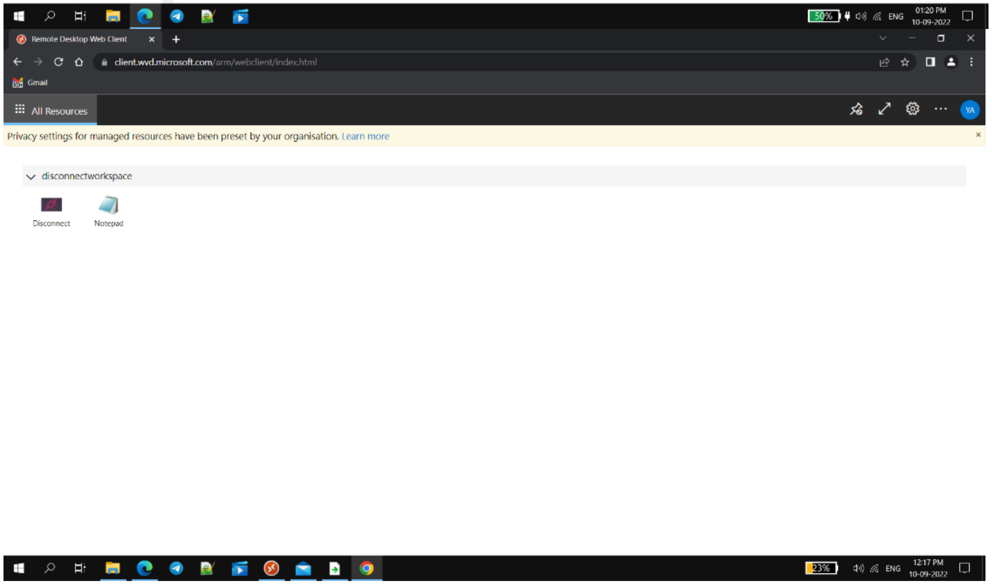
Security

Filter network traffic to and from Azure resources.

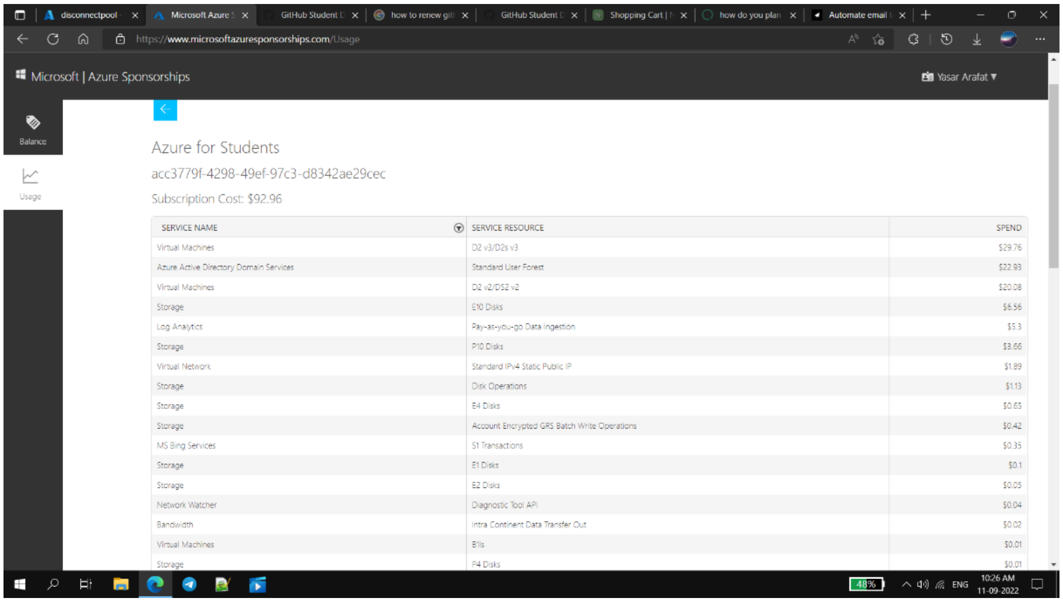
Private endpoints

Privately access Azure services without sending traffic across

● Not configured



Privacy settings for managed resources have been preset by your organisation. Learn more



Azure for Students

acc3779f-4298-49ef-97c3-d8342ae29cec

Subscription Cost: $92.96

SERVICE NAME	SERVICE RESOURCE	SPEND
Virtual Machines	D2 v3/D2s v3	$29.76
Azure Active Directory Domain Services	Standard User Forest	$22.93
Virtual Machines	D2 v2/DS2 v2	$20.08
Storage	E10 Disks	$6.56
Log Analytics	Pay-as-you-go Data Ingestion	$5.3
Storage	P10 Disks	$3.66
Virtual Network	Standard IPv4 Static Public IP	$1.89
Storage	Disk Operations	$1.13
Storage	E4 Disks	$0.65
Storage	Account Encrypted GRS Batch Write Operations	$0.42
MS Bing Services	S1 Transactions	$0.35
Storage	E1 Disks	$0.1
Storage	E2 Disks	$0.05
Network Watcher	Diagnostic Tool API	$0.04
Bandwidth	Intra Continent Data Transfer Out	$0.02
Virtual Machines	B1ls	$0.01
Storage	P4 Disks	$0.01

B. SOURCE CODE

Source code:

https://github.com/yasar-arafath/search

Search Engine web and app can be accessed on:

https://www.disconnect.cf