Module 1 - Data Acquisition & Programming

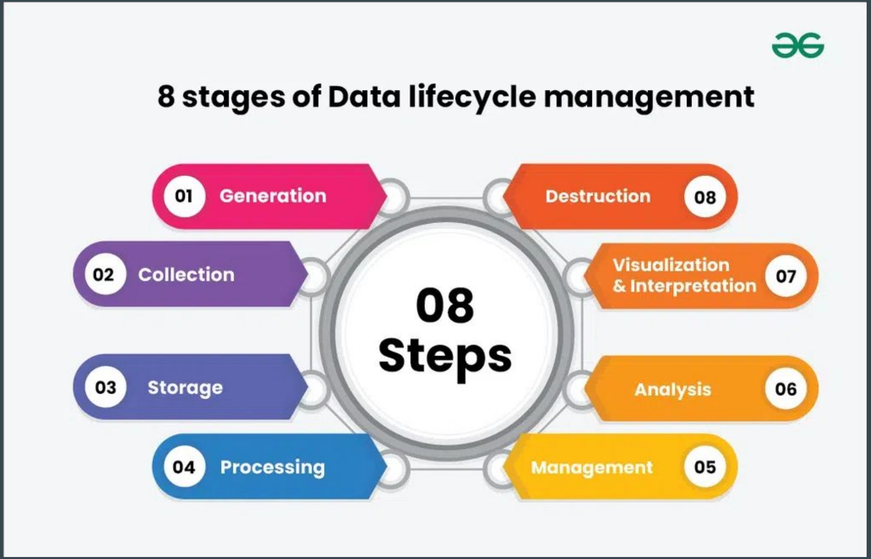
•••

CSc 46000 - Professor Sheng

Alex Abraham, Amir Nabiyev, Azim Rahat, Md Mamun, Wen Jie Long

What is Data Acquisition?

- Data acquisition is the process of gathering data from various sources for analysis
- Data can come from sensors, databases, the web, APIs, or real-time streams
- Critical for accurate insights and modeling

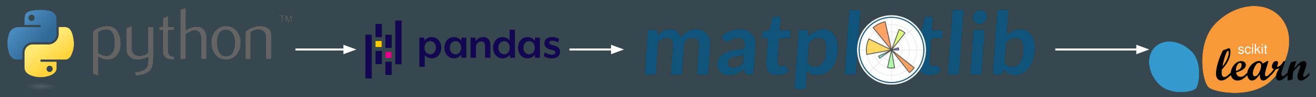


Data Science Programming Languages

- Languages used in data science are designed for efficiently handling large datasets, machine learning, and data visualization
- Most commonly used languages:
  - Python - Most widely known language
  - R - Designed for statistics and data visualization
  - MATLAB - Built for mathematical modeling and engineering

Python

- Python is an open-source, high-level language known for its readability and versatility
- Used in every step of the data science workflow
- Why it’s Popular
    - Extensive Libraries: NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, BeautifulSoup
    - Easy Integration: Works easily with web frameworks, databases, cloud services
    - Versatility: Used in analytics, AI research, automation, and engineering



Pull data from API

Clean data

Visualize data

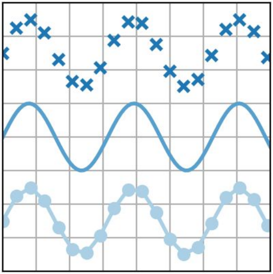
Train model

R

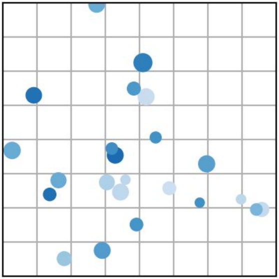
- R is an open source language built for statistical computing and graphs
- Useful in data analysis, hypothesis testing, and creating visualizations
- Why it's Popular
  - Easy Integration: Works with RStudio, able to interact with Python or SQL
  - Visualization Libraries: ggplot2 and plotly

MATLAB

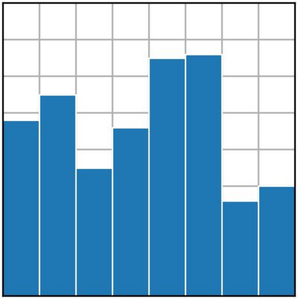
- A high-level language built for numerical computation, visualization, and algorithm development
- Why it's Popular
  - Matrix and math focus: Optimized for signal processing, control systems, and modeling
  - Built-in Toolkits: Specialized toolkits for image processing, machine learning, and statistics
  - Easy Integration: Easily connects with hardware sensors and Python for data collection



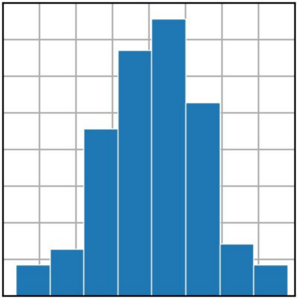
Plot



Scatter



Bar



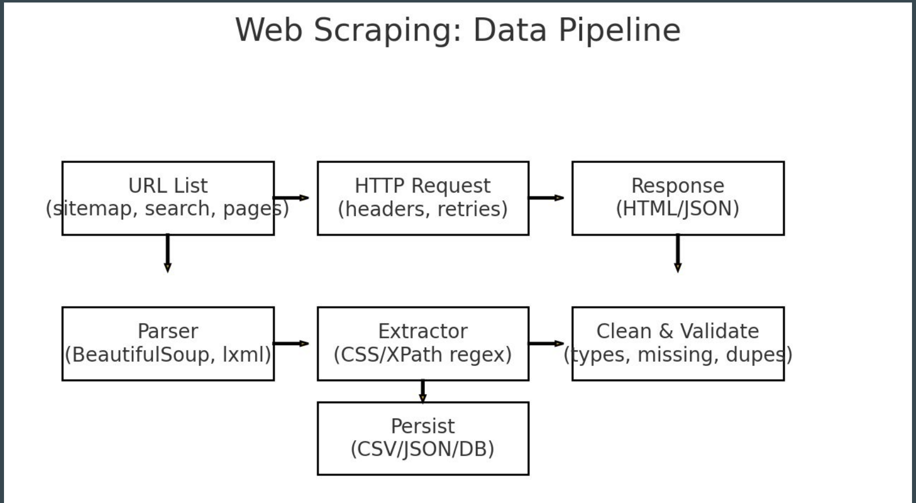
Histogram

What is Data Extraction in Data Science?

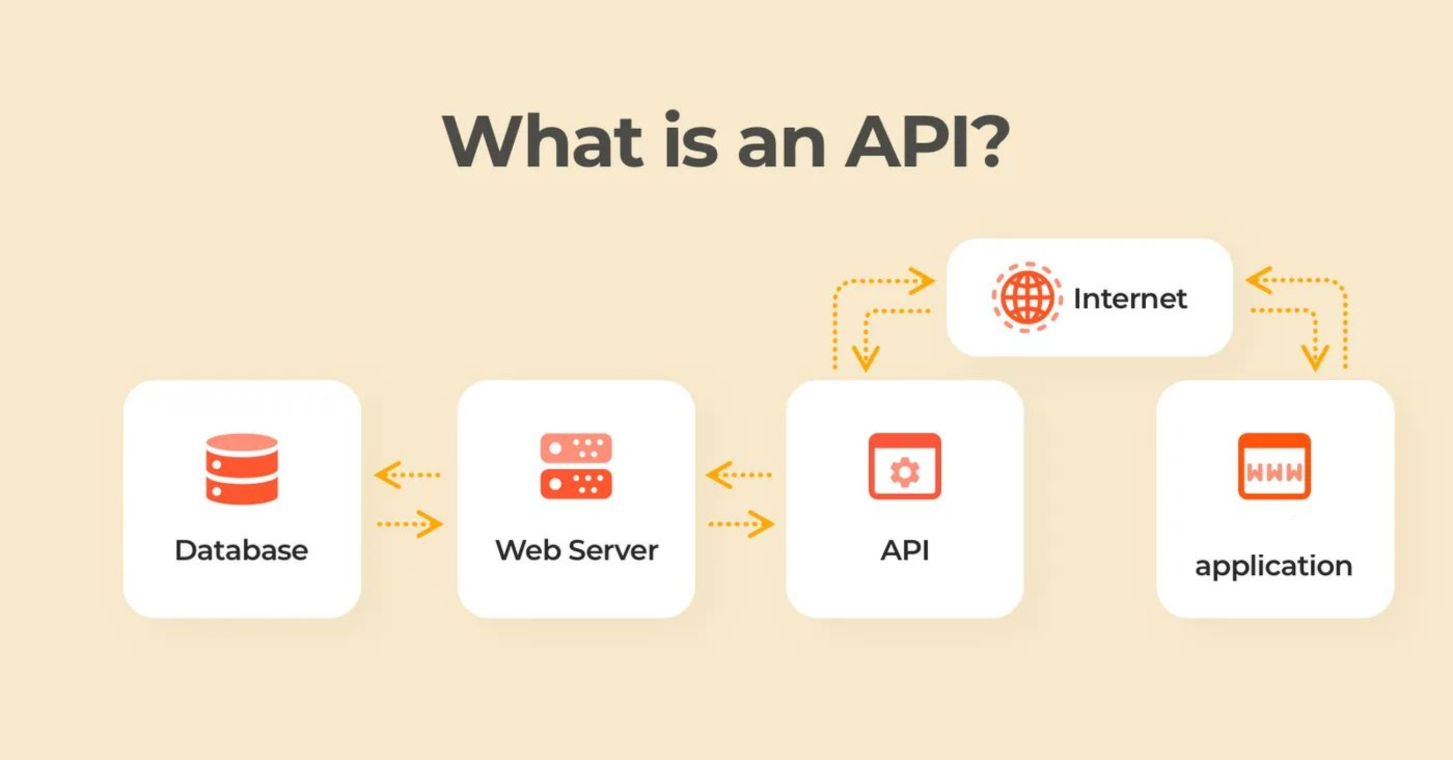
- Data extraction refers to to retrieving information from different sources including
  - Websites, databases, APIs, sensors, etc.
- Main techniques include
  - Web Scraping
  - APIs (Application Programming Interfaces)
  - cURL
  - WebSockets

Web Scraping

- The process of extracting information from webpages by parsing their HTML structure
- How it works
  - Send a request to the website's page
  - Retrieve the HTML source code
  - Parse and extract specific elements
  - Store the data in a set format (CSV, JSON, or database)


Web Scraping: Data Pipeline
URL List (sitemap, search, pages)
HTTP Request (headers, retries)
Response (HTML/JSON)
Parser (BeautifulSoup, lxml)
Extractor (CSS/XPath regex)
Clean & Validate (types, missing, dupes)
Persist (CSV/JSON/DB)

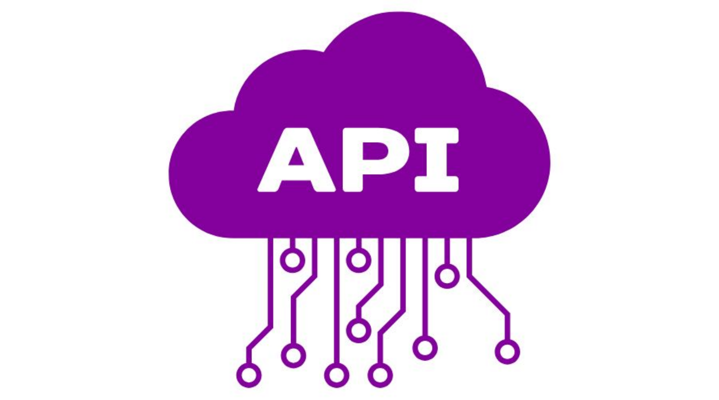
API - Application Programming Interface (Part 1)



https://www.cleveroad.com/blog/what-is-an-api/

API - Application Programming Interface (Part 2)

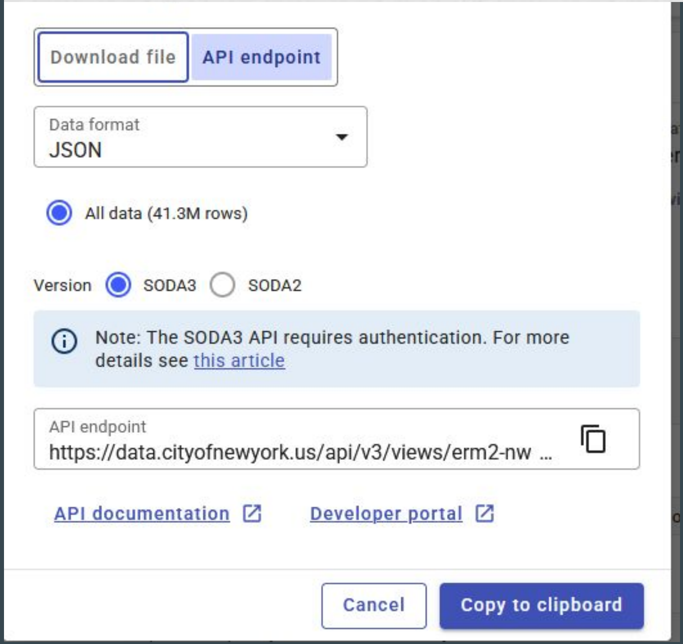
- An official method provided by a website to retrieve data from them
- Different sites have different APIs
- Can be free or paid
- Can have a rate limit
- Can require an API key as an identifier



API - Application Programming Interface (Part 3)

NYC OpenData

Websites with APIs will tell you how to use them.



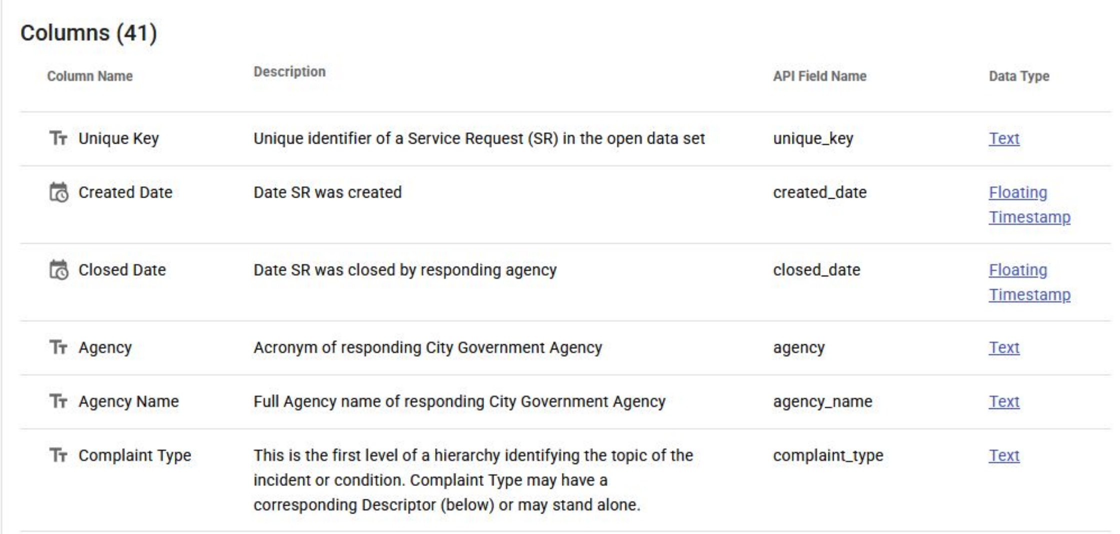
Data format
JSON

● All data (41.3M rows)

Version ● SODA3 ○ SODA2

Note: The SODA3 API requires authentication. For more details see this article

API endpoint
https://data.cityofnewyork.us/api/v3/views/erm2-nw ...

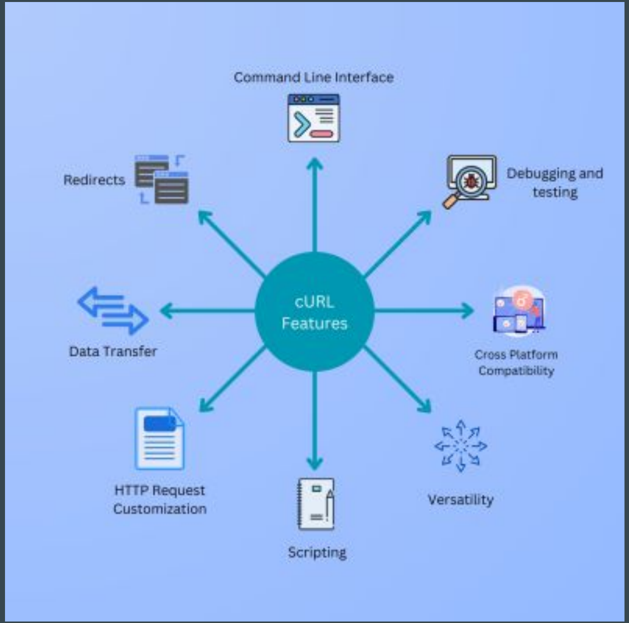


Columns (41)

Column Name	Description	API Field Name	Data Type
Unique Key	Unique identifier of a Service Request (SR) in the open data set	unique_key	Text
Created Date	Date SR was created	created_date	Floating Timestamp
Closed Date	Date SR was closed by responding agency	closed_date	Floating Timestamp
Agency	Acronym of responding City Government Agency	agency	Text
Agency Name	Full Agency name of responding City Government Agency	agency_name	Text
Complaint Type	This is the first level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.	complaint_type	Text

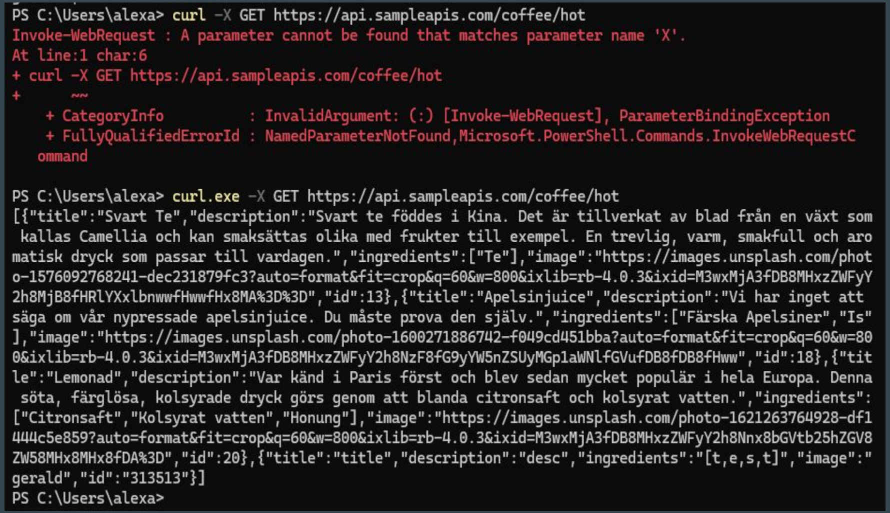
cURL (Client URL)

- Command-line utility for transferring data to and from a server
  - HTTP, HTTPS, FTP, SCP, and SFTP
- Usages:
  - Testing REST APIs (GET, POST, PUT, or DELETE)
  - Downloading files
  - Verify a website is running
  - Retrieve HTML content



cURL Syntax

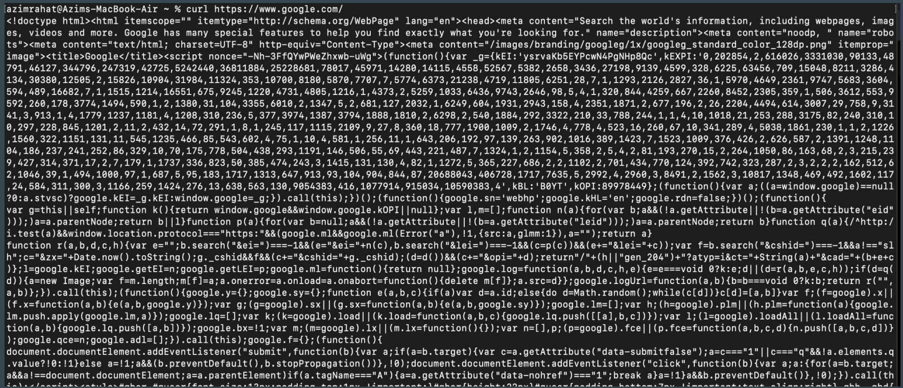
- curl [options] [URL]
- curl.exe [options] [URL]



```
PS C:\Users\alexa> curl -X GET https://api.sampleapis.com/coffee/hot
Invoke-WebRequest : A parameter cannot be found that matches parameter name 'X'.
At line:1 char:6
+ curl -X GET https://api.sampleapis.com/coffee/hot
+      ~~
    + CategoryInfo          : InvalidArgument: (:) [Invoke-WebRequest], ParameterBindingException
    + FullyQualifiedErrorId : NamedParameterNotFound,Microsoft.PowerShell.Commands.InvokeWebRequestC
   ommand

PS C:\Users\alexa> curl.exe -X GET https://api.sampleapis.com/coffee/hot
[{"title":"Svart Te","description":"Svart te föddes i Kina. Det är tillverkat av blad från en växt som
 kallas Camellia och kan smaksättas olika med frukter till exempel. En trevlig, varm, smakfull och aro
matisk dryck som passar till vardagen.","ingredients":["Te"],"image":"https://images.unsplash.com/phot
o-1576092768241-dec231879fc3?auto=format&fit=crop&q=60&w=800&ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxzZWFyY
2h8MjB8fHRlYXxlbnwwfHwwfHx8MA%3D%3D","id":13},{"title":"Apelsinjuice","description":"Vi har inget att
säga om vår nypressade apelsinjuice. Du måste prova den själv.","ingredients":["Färska Apelsiner","Is"
],"image":"https://images.unsplash.com/photo-1600271886742-f049cd451bba?auto=format&fit=crop&q=60&w=80
0&ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxzZWFyY2h8NzF8fG9yYW5nZSUyMGp1aWNlfGVufDB8fDB8fHww","id":18},{"tit
le":"Lemonad","description":"Var känd i Paris först och blev sedan mycket populär i hela Europa. Denna
 söta, färglösa, kolsyrade dryck görs genom att blanda citronsaft och kolsyrat vatten.","ingredients":
["Citronsaft","Kolsyrat vatten","Honung"],"image":"https://images.unsplash.com/photo-1621263764928-df1
444c5e859?auto=format&fit=crop&q=60&w=800&ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxzZWFyY2h8Nnx8bGVtb25hZGV8
ZW58MHx8MHx8fDA%3D","id":20},{"title":"title","description":"desc","ingredients":"[t,e,s,t]","image":"
gerald","id":"313513"}]
PS C:\Users\alexa>
```

curl -X GET https://api.sampleapis.com/coffee/hot

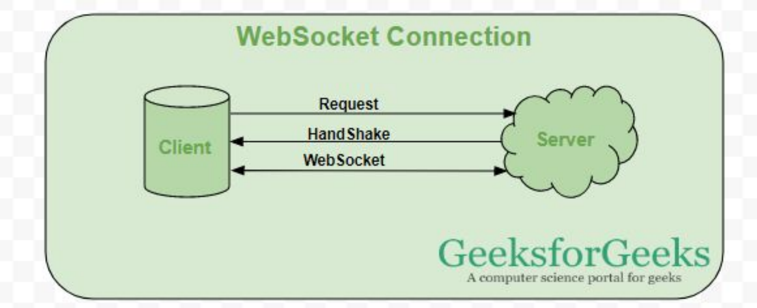


```
azimrahat@Azims-MacBook-Air ~ % curl https://www.google.com/
<!doctype html><html itemscope="" itemtype="http://schema.org/WebPage" lang="en"><head><meta content="Search the world's information, including webpages, imag
es, videos and more. Google has many special features to help you find exactly what you're looking for." name="description"><meta content="noodp, " name="robo
ts"><meta content="text/html; charset=UTF-8" http-equiv="Content-Type"><meta content="/images/branding/googleg/1x/googleg_standard_color_128dp.png" itemprop="
image"><title>Google</title><script nonce="-Nh-3FfQYwPWeZhxwb-uWg">(function(){var _g={kEI:'ysrvaKb5EYPcwN4PgNHp8Qc',kEXPI:'0,202854,2,616026,3331030,90133,48
791,46127,344796,247319,42725,5242440,36811884,25228681,78017,45971,14280,14115,4558,52567,5382,2658,3436,27198,9139,4599,328,6225,63456,709,15048,8211,3286,4
134,30380,12505,2,15826,10904,31984,11324,353,10700,8180,5870,7707,7,5774,6373,21238,4719,11805,6251,28,7,1,1293,2126,2827,36,1,5970,4649,2361,9747,5683,3604,
594,489,16682,7,1,1515,1214,16551,675,9245,1220,4731,4805,1216,1,4373,2,5259,1033,6436,9743,2646,98,5,4,1,320,844,4259,667,2260,8452,2305,359,1,506,3612,553,9
592,260,178,3774,1494,590,1,2,1380,31,104,3355,6010,2,1347,5,2,681,127,2032,1,6249,604,1931,2943,158,4,2351,1871,2,677,196,2,26,2204,4494,614,3007,29,758,9,31
41,3,913,1,4,1779,1237,1181,4,1208,310,236,5,377,3974,1387,3794,1888,1810,2,6298,2,540,1884,292,3322,210,33,788,244,1,1,4,10,1018,21,253,288,3175,82,240,310,1
0,297,228,845,1201,2,11,2,432,14,72,291,1,8,1,245,117,1115,2109,9,27,8,360,18,777,1900,1009,2,1746,4,778,4,523,16,260,67,10,341,289,4,5038,1861,230,1,1,2,1226
,1560,322,1151,131,11,545,1235,466,85,543,602,4,75,1,10,4,581,1,256,11,1,643,206,192,97,139,263,902,1016,389,1423,7,1523,1009,376,426,2,626,587,2,1391,1248,11
04,186,237,241,252,86,329,10,70,175,778,504,438,293,1191,146,506,55,69,443,221,487,7,1324,1,2,1154,5,358,2,5,4,2,81,193,270,15,2,264,1050,86,163,68,2,3,215,23
9,427,314,371,17,2,7,179,1,1737,336,823,50,385,474,243,3,1415,131,130,4,82,1,1272,5,365,227,686,2,2,1102,2,701,434,770,124,392,742,323,287,2,3,2,2,2,162,512,6
2,1046,39,1,494,1000,97,1,687,5,95,183,1717,1313,647,913,93,104,904,844,87,20688043,406728,1717,7635,5,2992,4,2960,3,8491,2,1562,3,10817,1348,469,492,1602,117
,24,584,311,300,3,1166,259,1424,276,13,638,563,130,9054383,416,1077914,915034,10590383,4',kBL:'B0YT',kOPI:89978449};(function(){var a;((a=window.google)==null
?0:a.stvsc)?google.kEI=_g.kEI:window.google=_g;}).call(this);})();(function(){google.sn='webhp';google.kHL='en';google.rdn=false;})();(function(){
var g=this||self;function k(){return window.google&&window.google.kOPI||null};var l,m=[];function n(a){for(var b;a&&(!a.getAttribute||!(b=a.getAttribute("eid"
)));)a=a.parentNode;return b||l}function p(a){for(var b=null;a&&(!a.getAttribute||!(b=a.getAttribute("leid")));)a=a.parentNode;return b}function q(a){/^http:/
i.test(a)&&window.location.protocol==="https:"&&(google.ml&&google.ml(Error("a"),!1,{src:a,glmm:1}),a="");return a}
function r(a,b,d,c,h){var e="";b.search("&ei=")===-1&&(e="&ei="+n(c),b.search("&lei=")===-1&&(c=p(c))&&(e+="&lei="+c));var f=b.search("&cshid=")===-1&&a!=="sl
h";c="&zx="+Date.now().toString();g._cshid&&f&&(c+="&cshid="+g._cshid);(d=d())&&(c+="&opi="+d);return"/"+(h||"gen_204")+"?atyp=i&ct="+String(a)+"&cad="+(b+e+c
)};l=google.kEI;google.getEI=n;google.getLEI=p;google.ml=function(){return null};google.log=function(a,b,d,c,h,e){e=e===void 0?k:e;d||(d=r(a,b,e,c,h));if(d=q(
d)){a=new Image;var f=m.length;m[f]=a;a.onerror=a.onload=a.onabort=function(){delete m[f]};a.src=d}};google.logUrl=function(a,b){b=b===void 0?k:b;return r("",
a,b)};}).call(this);(function(){google.y={};google.sy={};function e(a,b,c){if(a)var d=a.id;else{do d=Math.random();while(c[d])}c[d]=[a,b]}var f;(f=google).x||
(f.x=function(a,b){e(a,b,google.y)});var g;(g=google).sx||(g.sx=function(a,b){e(a,b,google.sy)});google.lm=[];var h;(h=google).plm||(h.plm=function(a){google.
lm.push.apply(google.lm,a)});google.lq=[];var k;(k=google).load||(k.load=function(a,b,c){google.lq.push([[a],b,c])});var l;(l=google).loadAll||(l.loadAll=func
tion(a,b){google.lq.push([a,b])});google.bx=!1;var m;(m=google).lx||(m.lx=function(){});var n=[],p;(p=google).fce||(p.fce=function(a,b,c,d){n.push([a,b,c,d])}
);google.qce=n;google.adl=[];}).call(this);google.f={};(function(){
document.documentElement.addEventListener("submit",function(b){var a;if(a=b.target){var c=a.getAttribute("data-submitfalse");a=c==="1"||c==="q"&&!a.elements.q
.value?!0:!1}else a=!1;a&&(b.preventDefault(),b.stopPropagation())},!0);document.documentElement.addEventListener("click",function(b){var a;a:{for(a=b.target;
a&&a!==document.documentElement;a=a.parentElement)if(a.tagName==="A"){a=a.getAttribute("data-nohref")==="1";break a}a=!1}a&&b.preventDefault()},!0);}).call(th
```

curl https://google.com

Websockets

- Communication protocol
  - HTTP is short-lives and simple
  - WebSocket real-time data exchange without repeated requests.
- Usage
  - Chat applications (WhatsApp, Messenger)
  - Collaborative apps (Google Docs)
  - Live dashboards or stock tickers (real-time updates)
  - Multiplayer games



- Connection Start – HTTP request with Upgrade
- Handshake – Server gives 101 Switching Protocols
- Persistent TCP Link – Switches from HTTP to WebSocket protocol, staying open.
- Close Handshake – One side sends a close frame; the other replies, then the TCP link ends.

Q&A