Omar AZZOUZI

Introduction

In the era of big data and fast-paced technological advancements, understanding and utilizing powerful data analysis tools are essential for future data engineers. One such tool is Apache Druid. In this report, I will provide an overview of Apache Druid, its applications, and why learning it is crucial for students interested in data engineering and real-time data processing.

What is Apache Druid?

Apache Druid, often referred to as just "Druid," is an open-source, high-performance, real-time analytics database designed for exploratory data analysis, business intelligence, and operational monitoring. It is built to manage and query large volumes of data in real-time, making it a valuable resource for businesses and organizations seeking to harness the power of data-driven decision-making.

Apache Druid is like a super-smart data warehouse for real-time data. It's designed to quickly store, search, and analyze large amounts of data that keeps coming in fast.

Applications of Apache Druid

1. Real-time Data Analysis:

Apache Druid excels at handling real-time data streams, making it an invaluable asset for applications requiring instant insights. With Druid, you can analyze and visualize data as it arrives, enabling businesses to make informed decisions promptly. This feature is crucial for various industries, including e-commerce, finance, and online advertising.

2. Interactive Analytics:

Druid's ability to provide sub-second query response times empowers users to interact with their data, ask ad-hoc questions, and explore trends and patterns with ease. This interactive analytics capability is vital for business analysts, data scientists, and decision-makers who require on-the-fly insights.

3. Historical Data Analysis:

In addition to real-time data, Druid can store and analyze historical data, allowing users to perform in-depth historical data analysis. This is particularly important for

understanding long-term trends and historical performance, crucial for industries like retail, where seasonal patterns play a significant role.

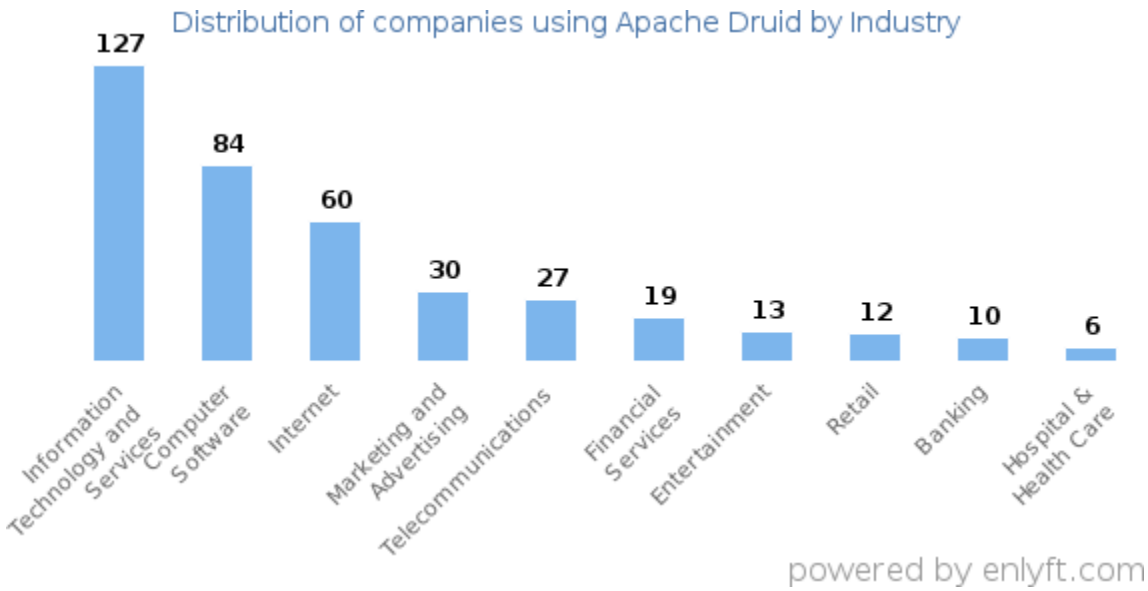
4. Data Exploration:

Druid's indexing and querying capabilities enable data exploration at a granular level. Students can utilize Druid to gain proficiency in discovering hidden insights within large datasets. It simplifies the process of filtering, aggregating, and visualizing data, enhancing data exploration skills.

Why Learn Apache Druid?

1. In-Demand Skills:

As the world becomes increasingly data-driven, organizations seek professionals with expertise in data analysis and real-time data processing. Learning Apache Druid equips data engineering with an in-demand skill set that enhances their employability.



2. Real-time Data:

Druid's real-time data processing capabilities are crucial in various sectors. Students who master Druid gain the ability to work with live data streams, making them valuable assets for organizations looking to maintain a competitive edge in the digital age.

3. Enhanced Data Analysis Skills:

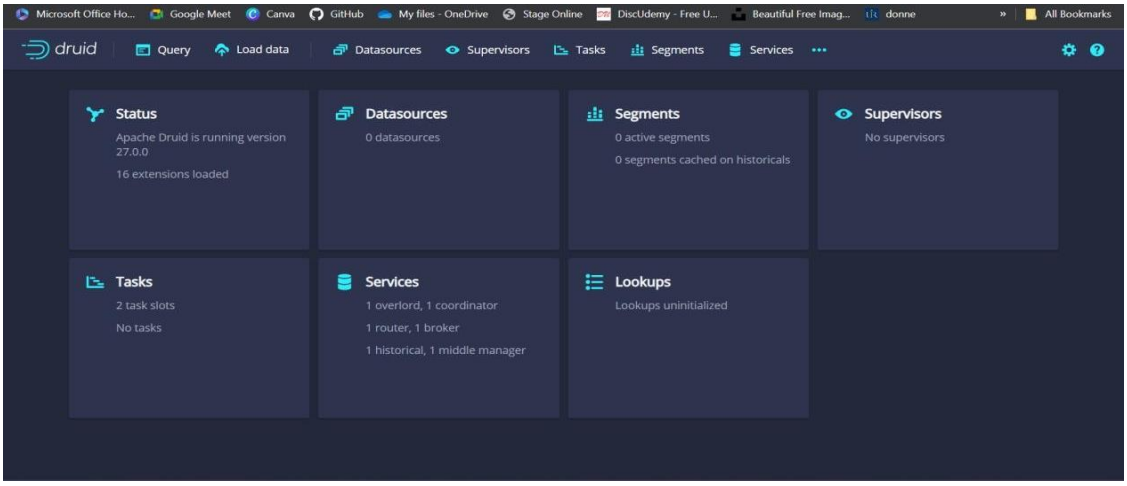
Apache Druid's querying and analytics features enable students to improve their data analysis skills. They can conduct more complex analyses, uncover insights faster, and make data-driven decisions with confidence.

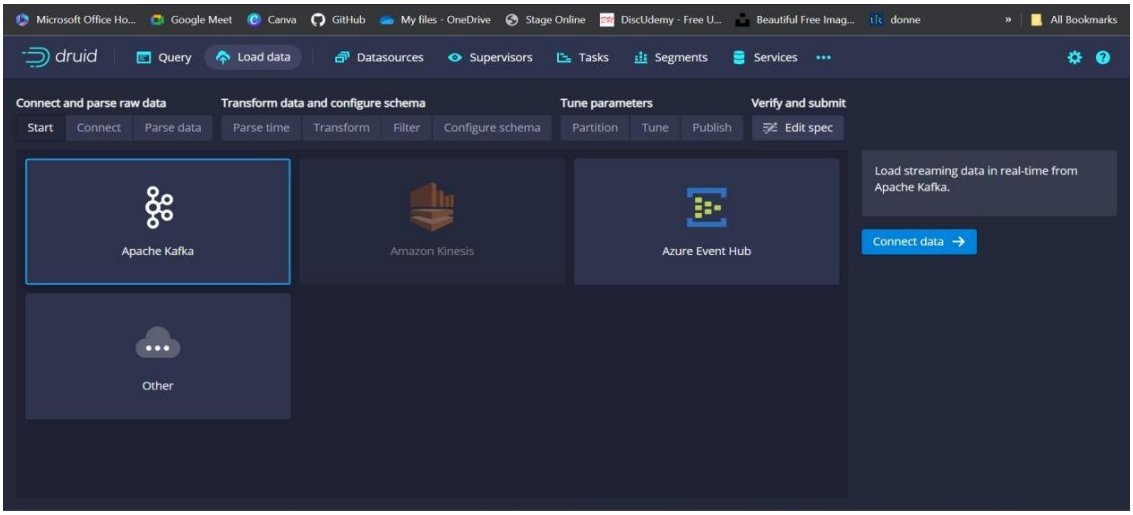
Project: Real-time Data Streaming and Analytics with Kafka and Druid

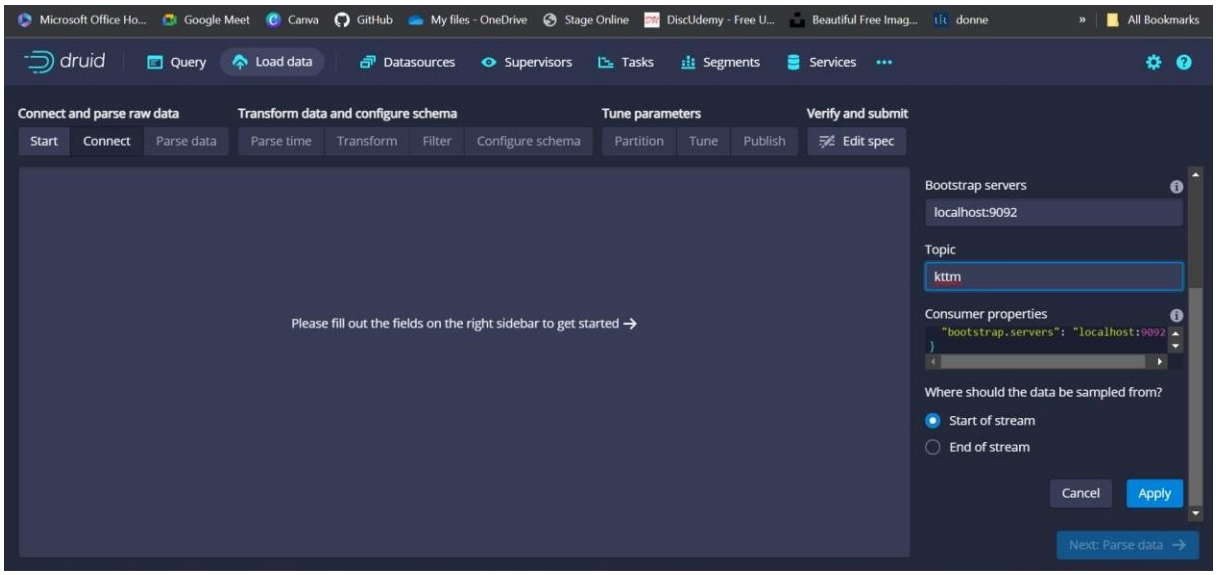
The project's scope encompassed setting up and configuring Apache Kafka and Apache Druid to create a data pipeline for ingesting, processing, and analyzing streaming data. I wanted to explore how these technologies could enable real-time analytics in various applications.

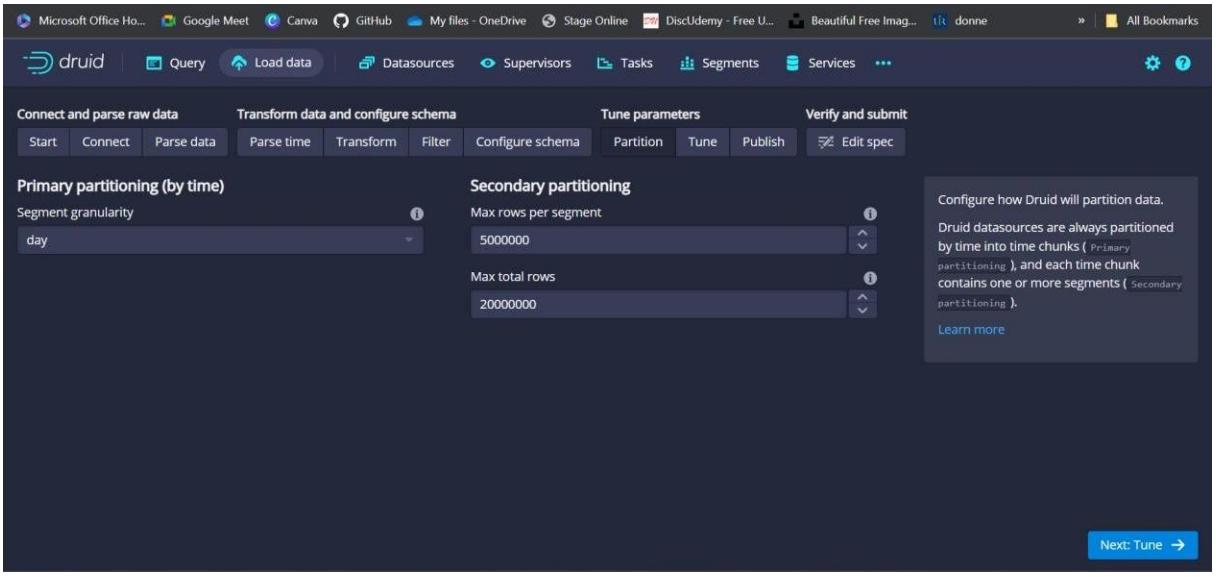
Project Implementation:

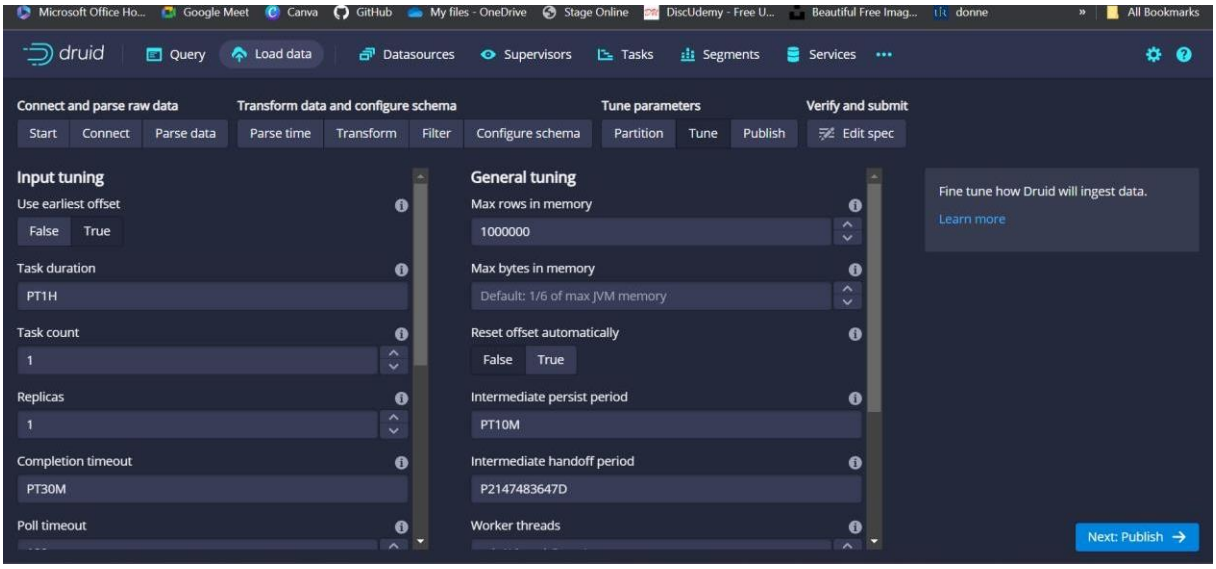
I started with the installation and configuration of both Kafka and Druid. While Kafka was employed to set up a data streaming platform, Druid was used to store and analyze the data. This involved downloading and setting up the necessary software components

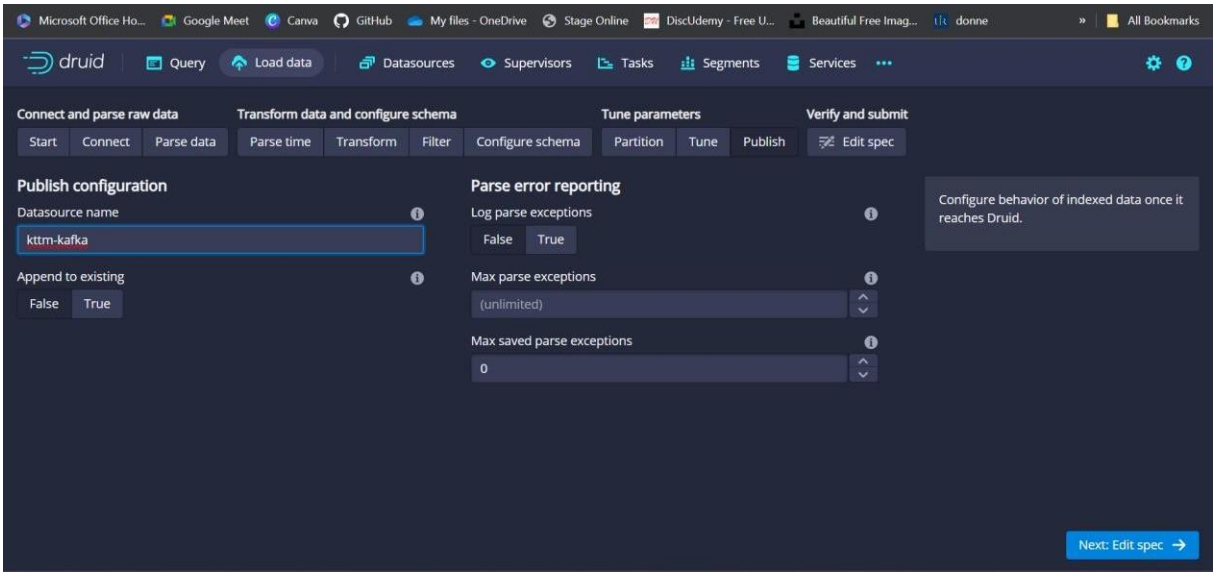


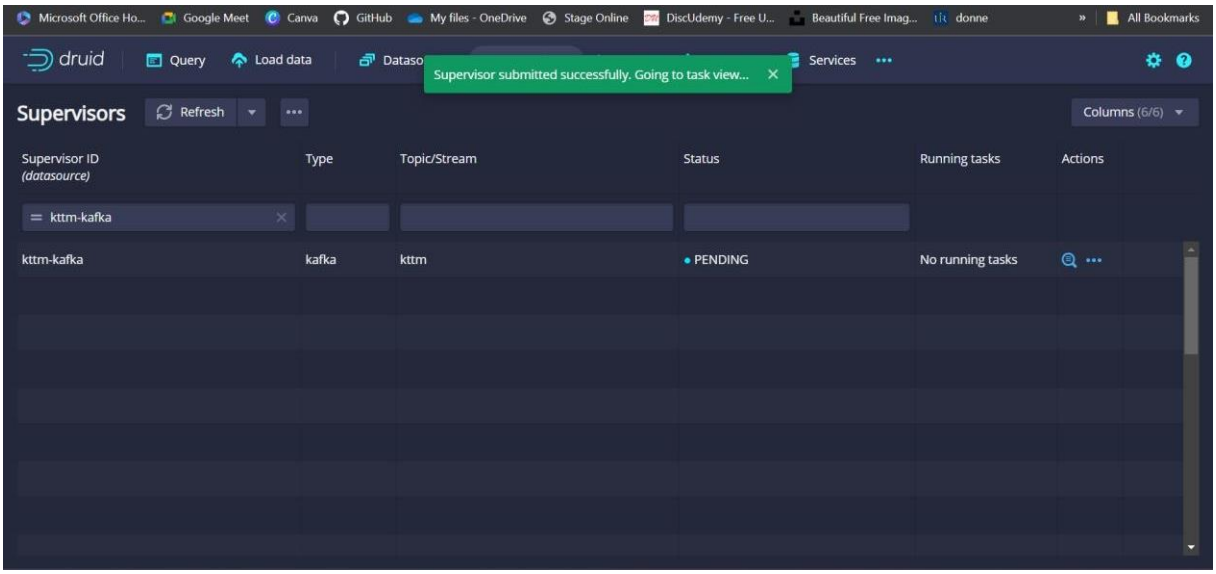



Microsoft Office Ho...
Google Meet
Canva
GitHub
My files - OneDrive
Stage Online
DiscUdemy - Free U...
Beautiful Free Imag...
donne
All Bookmarks
druid
Query
Load data
Datasources
Supervisors
Tasks
Segments
Services
Connect and parse raw data
Start
Connect
Parse data
Transform data and configure schema
Parse time
Transform
Filter
Configure schema
Tune parameters
Partition
Tune
Publish
Verify and submit
Edit spec
Please fill out the fields on the right sidebar to get started →
Bootstrap servers
localhost:9092
Topic
kttm
Consumer properties
"bootstrap.servers": "localhost:9092
}
Where should the data be sampled from?
Start of stream
End of stream
Cancel
Apply
Next: Parse data →

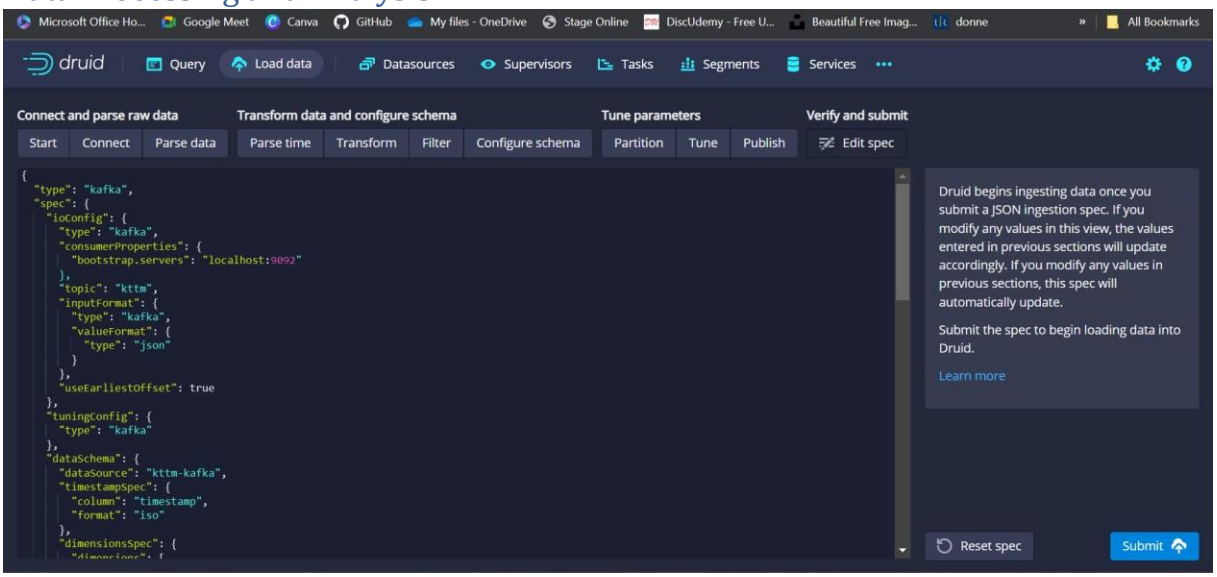

Microsoft Office Ho...
Google Meet
Canva
GitHub
My files - OneDrive
Stage Online
DiscUdemy - Free U...
Beautiful Free Imag...
donne
All Bookmarks
druid
Query
Load data
Datasources
Supervisors
Tasks
Segments
Services
Connect and parse raw data
Start
Connect
Parse data
Transform data and configure schema
Parse time
Transform
Filter
Configure schema
Tune parameters
Partition
Tune
Publish
Verify and submit
Edit spec
Primary partitioning (by time)
Segment granularity
day
Secondary partitioning
Max rows per segment
5000000
Max total rows
20000000
Configure how Druid will partition data.
Druid datasources are always partitioned by time into time chunks (Primary partitioning), and each time chunk contains one or more segments (Secondary partitioning).
Learn more
Next: Tune →


Microsoft Office Ho... Google Meet Canva GitHub My files - OneDrive Stage Online DiscUdemy - Free U... Beautiful Free Imag... donne All Bookmarks
druid
Query
Load data
Datasources
Supervisors
Tasks
Segments
Services
Connect and parse raw data
Start
Connect
Parse data
Transform data and configure schema
Parse time
Transform
Filter
Configure schema
Tune parameters
Partition
Tune
Publish
Verify and submit
Edit spec
Input tuning
Use earliest offset
False
True
Task duration
PT1H
Task count
1
Replicas
1
Completion timeout
PT30M
Poll timeout
General tuning
Max rows in memory
1000000
Max bytes in memory
Default: 1/6 of max JVM memory
Reset offset automatically
False
True
Intermediate persist period
PT10M
Intermediate handoff period
P2147483647D
Worker threads
Fine tune how Druid will ingest data.
Learn more
Next: Publish


Microsoft Office Ho... Google Meet Canva GitHub My files - OneDrive Stage Online DiscUdemy - Free U... Beautiful Free Imag... donne All Bookmarks
druid
Query
Load data
Datasources
Supervisors
Tasks
Segments
Services
Connect and parse raw data
Start
Connect
Parse data
Transform data and configure schema
Parse time
Transform
Filter
Configure schema
Tune parameters
Partition
Tune
Publish
Verify and submit
Edit spec
Publish configuration
Datasource name
kttm-kafka
Append to existing
False
True
Parse error reporting
Log parse exceptions
False
True
Max parse exceptions
(unlimited)
Max saved parse exceptions
0
Configure behavior of indexed data once it reaches Druid.
Next: Edit spec

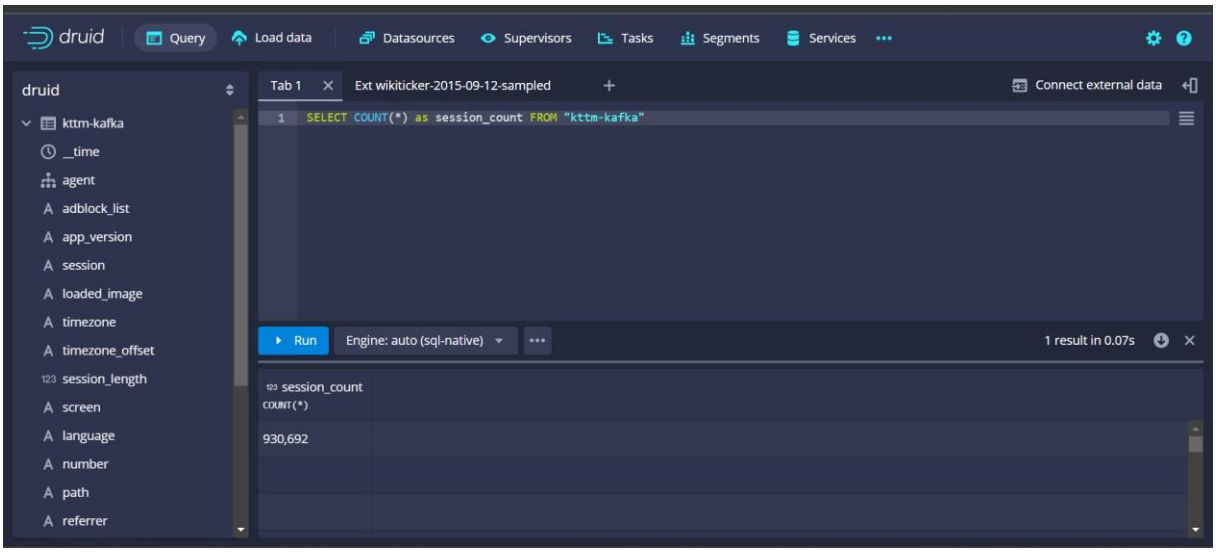


Data Processing and Analysis:



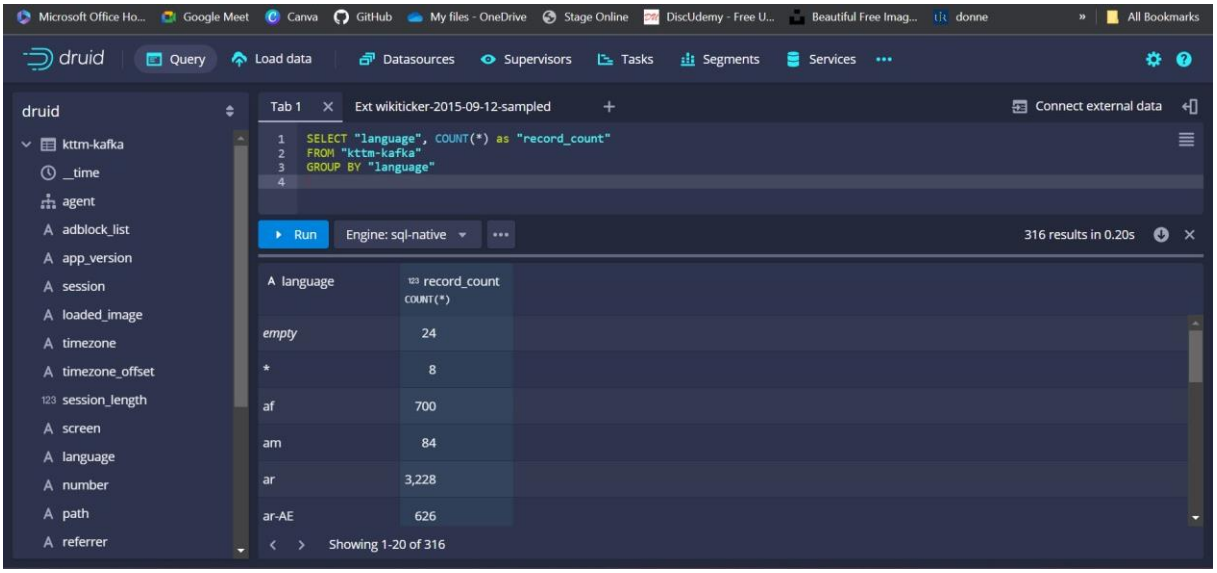
Once the data was ingested into Kafka, it was ready for consumption by Druid. This step involved configuring Druid to pull data from Kafka topics. I performed various data analytics by writing SQL queries to retrieve real-time insights from the streaming data. For instance, I queried the data to extract specific information, and the results provided insights in real time.

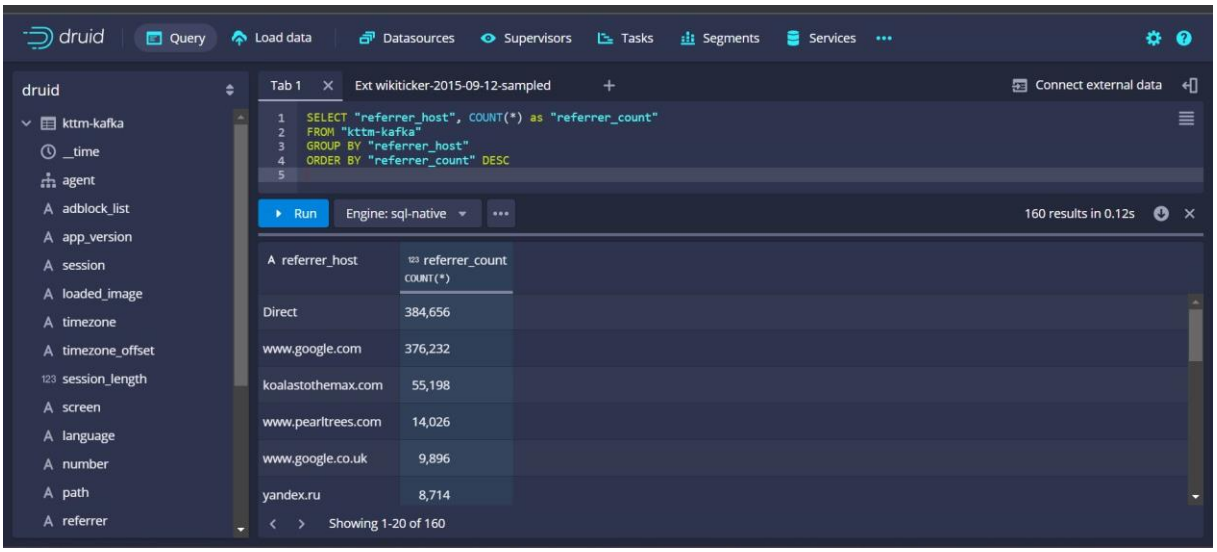
Count the number of sessions:



Count of Records by Language:

Top Referrer Hosts:




druid
Query
Load data
Datasources
Supervisors
Tasks
Segments
Services
druid
kttm-kafka
__time
agent
adblock_list
app_version
session
loaded_image
timezone
timezone_offset
session_length
screen
language
number
path
referrer
Tab 1
Ext wikiticker-2015-09-12-sampled
Connect external data
SELECT "referrer_host", COUNT(*) as "referrer_count"
FROM "kttm-kafka"
GROUP BY "referrer_host"
ORDER BY "referrer_count" DESC
Run
Engine: sql-native
160 results in 0.12s
referrer_host
referrer_count
COUNT(*)
Direct 384,656
www.google.com 376,232
koalastothemax.com 55,198
www.pearltrees.com 14,026
www.google.co.uk 9,896
yandex.ru 8,714
Showing 1-20 of 160