TAYLOR'S UNIVERSITY

Wisdom · Integrity · Excellence

BACHELOR IN COMPUTER SCIENCE (HONORS)

MAY 2024 – SEMESTER 5

MODULE CODE: ITS 66904

MODULE NAME: BIG DATA TECHNOLOGIES

GROUP (5) PROJECT ASSIGNMENT TASK 2 (30%)

Due Date: Saturday, July 27, 2024 via MyTIMeS (22:59PM Nepal Time)

Student Declaration: We Declare That –

- *we confirm the awareness about the university's regulations, governing cheating in tests and assignments, and form the guidance issued by the school of computing and it concerning plagiarism and proper academic practices, and the assessed work now submitted is in accordance with this regulation and guidance.*
- *we understand that, unless already agreed with the school of computing and it, that the assessed work has not been previously submitted, either in whole or in part, in this or any other institution.*
- *we recognize that should evidence emerge that my work fails to comply with either of the above declarations, then we may be liable to proceeding under regulation.*

Student Full Name	University ID	Signatures	Scores
Amogh Shakya	036 2073		
Monalisha Thapa Magar	036 2948		
Prajwol Dahal Khatri	036 2164		
Sujal Ratna Tuladhar	036 2483		
Xenium Christina Thebe	036 2911		

Table of Contents

Table of Figures

Section A:

1. Problem Analysis & Problem Requirement

for Netflix to maintain its lead and competitive edge to the current streaming industry, after being on the top for so long and now due to saturation of markets. and a vast number of companies creating their own streaming platforms and pay per view subscribers. These huge mega-corporates are in a continuous battle to get engagement from a diverse global audience to get their money monthly or yearly on deals that may entice them. so, they have to be involved and adapt to growing needs of addressing issues relating to content creation, preferences of the viewer and general audience, without overloading and over saturating the market. They have to decide on such core problems like what source material and type of content to acquire, produce, and recommend to retain existing subscribers and attract new ones. Since Netflix is driven by data since the beginning, the problem is compounding each year to the globalization of and population connected to the internet. so, to keep up with trends accurately. They have been experimenting on creating sophisticated analysis tools for their platform.

the need of robust strategy and leveraging advanced techniques to address the challenge and requirements which include analyzing large volumes of user generated data to discern patterns for effective recommendation, testing, to gauge the impact of different strategies employed from time to time. with such global reach across 190 countries and hundreds and thousands of users monthly. They are figuring out the balance of global with local, either with the help of data analytics, machine learning, recommendation algorithm, content making that predicts and relates and cater to diverse audiences and individual viewer's interest. running these huge applications 24-7-365 has a huge toll on maintaining resilience of the system-services and change-failures. These real-time content machines need to handle disruptive and failure scenarios effectively. it could cost millions to the company when it fails so it requires implementation of load shedding, fallback to remain operational and prevent overload

1.1. How Netflix Leverage Big Data Analytics

By harnessing the power of big data analytics, Netflix can enhance the content related strategy, improving user experience and propelling the decision-making process. the collection of various amounts of data related to behavior tracking, pattern viewing, querying and searching, interaction and ratings to gain insights to all minimal patterns to personalize their content recommendations algorithm to increase retention values. These contents are personalized to ensure that viewers have their interests and spend more time binge watching. by analyzing the trends, they can be ahead of the game of identifying emerging genres, popular themes, and high demand. Acquiring ability to make a data-driven and make sound decision to produce an original content or licensing adaption of external/existing.

Fast content delivery and distribution is also an important role in optimization and increasing the delivery of content, those seamless, and low latency to enhance viewers satisfaction and being chosen over other dynamic competition. such is also true for other various business dimensions, like merchandise application platform to apply business logic to recommendation "My List" and "Continue Watching" in real time across multiple devices. Here the historical data uses predictive models to analyze preferences, metrics.

1.2. Enhance Customer Experience

Customer experience is enhanced through a combination of content recommendations, user friendly interface, and integration across various forms of devices. These real time data processing helps in seamless transition between devices without losing responsiveness to enhance viewing experiences. this synchronization tracking help content playback

Personalized Recommendations: the users viewing history data, rating, search query, will continuously adapt based on interactions, the Netflix algorithm (collaborative and content-based filtering) will predict what they will enjoy and tailor-made suggestions. doing such the likelihood of engagement will resonate with watch time. **Content Strategy:** information of demographic, engagement metrics, viewing pattern and frequency, they can determine the respective resonating content. and testing can be done to increase engagement, user experience, content acquisition and production and marketing. **Streaming Quality:** dependent on user's device capabilities and internet speed. The stream quality is dynamically optimized with adaptive bit rates streaming. This ensures that buffering is minimized with low interruptions for pleasantness. **User Interface and experience:** designing is a huge part of Netflix, it is important due to seamless navigation, easy intuitive search functions and displaying content. across various devices and platforms, and resolutions and still be high quality. **Content library:** across 190 countries and many languages, Netflix has a diverse library, with various genres and formats. this idea ensures there is something for everyone on top of exclusive original content

1.3. Recommendation Engine

Data Collection: Netflix collects information like viewing history (what is watched, when it's watched and how long they have watched), the feedback is noted with ratings and reviews, search queries of titles, watch list and other types of similarity, networking conditions and device information.

Data Processing: combination of models **Collaborative Filtering Technique** uses matrix factorization methods like Singular Value Decomposition (SVD) to predict capture pattern of characteristics like what other users may and might enjoy watching based on their past history and user preferences, here the features are decomposed to user-item matrix factors. **Content-Based Filtering Technique** also based on previously liked contents, it analyzes the attributes inside the content like genres, actors, directors, studios and matches them. **Deep Learning** networks are used in analysis of complex patterns with deep embedding for accurate predictions.

A/B Testing and Optimization: Netflix are continuously testing with different algorithms and versions which are sent to the users to measure each performance like engagement, click rates, and satisfactions. such engines are able to adapt and stay relevant with change in behavior. **Feedback Integration**: implicit behavior (like skipping parts or watching to completion) and explicit ratings.

1.4. Content Strategy and Creation Decisions

using user behaviors and interactions metrics, completion rates, watching history, rewatches and search queries. Here the patterns, trends, viewing times, interests are guiding for content strategy, the plot, actors, or genre are gleaned from the data and will interest viewers. viewing past patterns, the choice of content to modify its strategy and use of highly advanced algorithms. and the materials align with what the users are interested in. here the frequent runs of A/B tests during content offerings that impact user engagement and moods and have higher retentions. Tracking performance is such a rigorous task. where high engagement indicates the effectiveness of content strategy implemented. by predicting the next possible popularity of new materials which may be chosen to fund and finance/produce any original content programs or movies. Here, based on discovering relevance and predicted success in the current market trends, it is also shown to worth or acquire content. with preference of current user and possibility to attract new members by making accessibility to watch content. based on geolocation and regional, contents are acquired to cater to a diverse global audience. investing in localized content markets, understanding and making sure that it doesn't offend their cultures and are respectful and sensible. by adapting frequency their new technologies are resilient.

As any top company would, Netflix also tracks and benchmark competitors' content offerings to discover gaps and opportunities, similarly as SWOT analysis. This aids in differentiating its content and developing distinct value propositions influenced by global entertainment trends with technologies. the proliferation of international content and a variety of genres in the library. uniqueness in making and providing originals series and documentations, documentaries that will set apart from rivals. Collaboration, investments, partnership and working together with talented and high-profile well-known filmmakers, performers, and producers is essential to creating compelling and high-caliber content for guaranteed and successful projects also not forgetting about bonding with local unique languages with regional flavors. helping in spreading, preserving and cementing their culture. boosting the subscription growth.

analyzing genres, actors, or directors that attract the most engagement. algorithms to determine which content would resonate with which portion of its audience. The use of machine learning models and multi label text classification help to refocus on genres and predict for Netflix. being on the lookout for a balance and financial assessment of the potential returns on investment. looking at the success rates, building to measure viewership and prospective revenue that will amass. The platform designs its content strategy in order to hit these different geographic markets. This data-driven approach is also coupled with finance strategic assessments and considerations on global markets to increase relevance and profitability for the content.

2. Initiative Driver

2.1. Specific Data Points

2.1.1. Viewing History:

Title Watched: keeping track and record of the titles that a customer watches, helping in listing recommendation and suggested content. **Viewing Duration:** amount of time spent watching (finished or stopped mid way), calculating preference and engagement level for each title in a specific genre provides insights into content more appealing to individual users. **Viewing Frequency:** knowing the habits (binge-watching patterns) and viewing patterns (how long, what time) of a user for content visibility. **Playing Interaction:** can also play part in refining the recommendations and better engagement.

2.1.2. User Interface Data:

Search Queries: fine tuning the search functionality and accurate mapping to content, when there are minor mistakes in the search bar, understanding the keywords. what they search. **Ratings and Thumbs Up/Down:** feeds the recommendation algorithm and ensures it is closely with their tastes. **Browsing Behavior**: clicking through the categories in the interface, point of pause or stopping **Content Preferences**: in essence with past viewing data, the ratings and likes on favorite genres, actors, themes that assists in improving on user satisfaction and interactions. **Feedback Loop** is utilized for creators to fine tune and meeting audience expectations. **Social Sharing and Media Sentiments.**

2.1.1. Contextual Data

Device Information: keeping track of types of devices in use to access the site, such as smart TVs, smart mobile devices, and desk/laptops. allowing for optimized content delivery and smooth performance **Location**: geographic location information is used to recommend based on the region, such help in correctly navigating and networking, adjusting preference and availability. **Time**: by determining the time of the day where large influx of users view content helping in modeling and determine peak times and make recommendations and extra secure and fail proof. as users may view lighter content during the evenings and more engaging or intense content during weekends and midnight. **Metadata** give additional information which is the easiest principle. **Demographic** like Age, Gender in tailoring and proper segmentations.

3. Technologies Employed by Netflix

3.1. Open Connect Netflix

providing millions of netflix subscribers with the highest quality viewing experience possible, by partnering with thousands od Internet Service Provider (ISPs) by localizing substancial amount of traffic with embedded application deployment. Such proprietary technology will cache contents on server those are statigically placed around the world to reduce latency with better streaming quality as it is accessing from more closer locations. especially during peakness to have smooth playback

3.2. Web Service

by relying heavily on best-in-class Amazon Web Service (AWS) for scalable cloud infrastructure, thousands of server/computing resources, terabytes of storage solutions, and database needs that are vital for handling Netflix's vast amount of data, analytic recommendations, video transcoding, and streaming requirements. This flexibility and scalability of AWS services allow Netflix and top artistic talents to efficiently and quickly deploy manage traffic spikes and ensure uninterrupted service delivery functions using 100,000 server instances. for over 260 million members in 190 countries, 30 languages, 125 million watch hours entertaining the world

3.1. Amazon S3

use of hybrid cloud storage that is the combination of the AWS and Simple Storage Services (S3) for scalable and reliable, for durability in vast data amounts. with 'Local Zones' tracking enormous assets with multipart object stores, access control, user data, logs, back up, life cycle managements, crucial for extensive data requirements, in decentralized storage and centralized management control.

3.2. Hadoop

being open source distributed storage for large dataset processing across clusters of computers. which is vitally used by netflix to handle massive amount of data everyday including warehouses and batch processing, having ability to scale horizontally (across multitude of serveers) they are able to process across large volumes effectively/efficiently

3.3. Hive

a central warehousing insfrastructure built on top of hadoop that enables simpliefied querying and managing gargantuan datasets, with translated SQL like interface and quries into MapReduce jobs, enabling users with convenient way to perform complex data analysis without low-level code. in its extensive data repositories, facilitating insights and metrics. extracting actionable insights.

3.4. Pig

a high-level platform for creating MapReduce programs that used on Hadoop. a scripting language, Pig Latin, which simplifies/abstracts the complexity of writing MR2 code (usally raw low level) used for large scale data processing data transformations and aggregations, allowing efficiency to data engineers.

3.5. HBase

a NoSQL database that provides real-time read.write access to large volume datasets and operates on top of Hadoop. due to its availability and scalable design. dataser are able to be managed and stored data in a distributed manner, for on the spot analytics and large-scale processing.

3.6. Kafka and Spark

integral for real-time data streaming pipeline, processing in large volume and minimal delay helpful in decision making process by tracking behaviors in real-time while ingesting trillions of events accurately.

used in. running complex parallel data processing, analytics and machine learning algorithms, which is enabled quickly due to in-memory computing and efficiency. with falut tolerance.

3.7. Tensor Flow and PyTorch

a framework for deep learning to construct machine learning models like powering advanced recommendation, predictive algorithms, content classification. and many more due to it ML research domain that optimize and improve engagement

3.8. Java and Impala

the core fundamental programming language of backend systems. as its reliability and performance, various components infrastructure were built on **Java**, especially microservices and processing apps. due to it robust ecosystem, high performance, scalable and supported by community developing idela concurrent programming.

Impala an open source engine for SQL that is interactive and fast for use of hadoop. being able to run complex queires in faster quicky in real-time,

3.9. Cassandra and DynamoDB

Cassandra distributed NoSQL database which designe to handle volumns with more scalability is very available to manages large data across multiple nodes, ensuring guaranteed consistentency, in various applications, managing user, supporting high-traffic services, overcoming and benefiting with global opeation.

DynamoDB is a fast and flexible that complements Netflix need for managing real-time data and backing scalable in term of high-speed access and applications.

3.10. Ffmpeg/VMAF

An open-source, cross-platform, Emmy-winning library used in a huge number of metric multimedia applications, combining human vision and ML/NN. It efficiently optimizing encodes (AV1 and leverage spatial and temporal redundancies) and compresses digital video, using Neural Network working on the perfecting of video quality, investing on next-gen, royalty free codecs with regard to bandwidth consumption, so that content can be delivered to its end-users in the most efficient way over different devices and network conditions. downscaling in preprocessing, with also multiple resolutions resampling filters.

3.11. HTML5 and Adaptive Bitrate Streaming

by moving into standardized video playback technologies enabling flow of video content across all browsers and cross-platform compatibility of electronic devices. switching between video qualities dynamically based on the network connections, changing in real time continuously do the technology like HLS and MPEG-DASH helps in not breaking, pausing/buffering.

3.12. Deployment with Monitoring

Docker containers are used in Netflix apps. deploying services in insolated manageable environments that are consistent across the web. **Kubernetes** clusters also orchestrate as managing microservices, and large deployment. monitoring is super systematic process of aggerating actionable metrics and logs, using **Prometheus** (open source) performance alerting toolkit and **Grafana** to visualize and make dashboard easily accessible and quick to spot problem

3.13. Security

TLS (Transport Layer Security) and SSL (Secure Socket Layer) Encryption is to ensure/guarantee that the information's are safely exchanged privately between company servers and user devices to prevent data leakage. **AWS Shield** gives protection against DDoS attacks, protecting the infrastructure against cyber-attacks.

3.14. Back End and API

Spring Boot is a large scale, robust backend Java-based framework with many settings and library components. meant to install, develop, build, deploy solid services that are vital to different services that will communicate effectively even if it works independently in rich ecosystem making backend scaling possible.

Graph-QL is used as advanced query manipulation language that helps requests sent by clients to point exact data required, no more or less. fetching data has made it efficient and flexible to interact in comparisons with traditional RESTful. these precision and reduced loads and data transferred over a network improve the performance with user experience

4. Diagrams

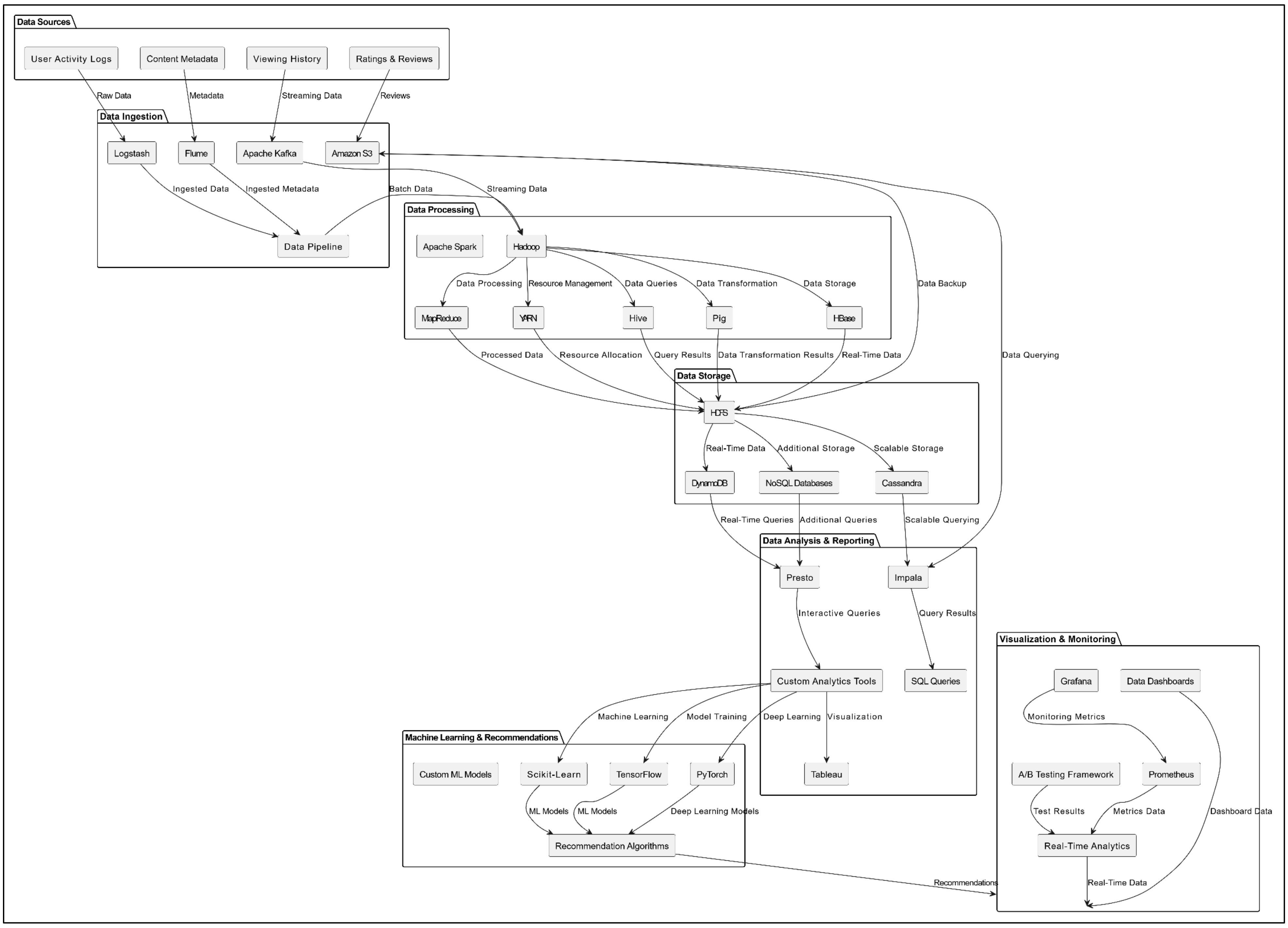


Figure 1. Netflix Infrastructure

5. Netflix's Reliance on Big Data

5.1. Influence on Evolution

Netflix has had a great deal of influence on how big data technologies and practices have been developed. Being one of the largest streaming platforms in the world, several developments of infrastructural aspects.and innovational trends resulted from the ways the company dealt with big data. The large scale of Netflix forced the organization to look for technologies like from Apache setting a standard for real-time data storaging, streaming and large-scale data processing. This impacted other organizations into emulating it in dealing with their big data challenges. On the basis of TensorFlow and PyTorch, the popularized recommendation algorithm of Netflix creates new standards for the digital and physical user experience. improving the solution of content delivery by caching the content, the potential of enhancing the quality of streaming by decreasing latency, encouraging all to invest in similar platforms. Moreover, tools that Netflix used, also not only improved but have also encouraged other organizations towards the technologies. In some cases, thanks to data processing technologies, several development have appeared with the frameworks of data processing. For example, using Kafka will focus on adopting event-driven paradigms as a common practice within for the continuous flow of data. Additionally, the machine learning libraries and employing has improved results have set industry benchmarks that other corporations strive for. the production of the open connection and Netflix's focus on good and efficient streams continuously pushed the boundaries of content delivery and streaming. Further, uses of very extensive cloud services, big data tools and technologies, unveiled new characteristics of the big data infrastructure and scalability. fixation with security, other organizations also follow the same with an aim of protecting their data and their privacy.

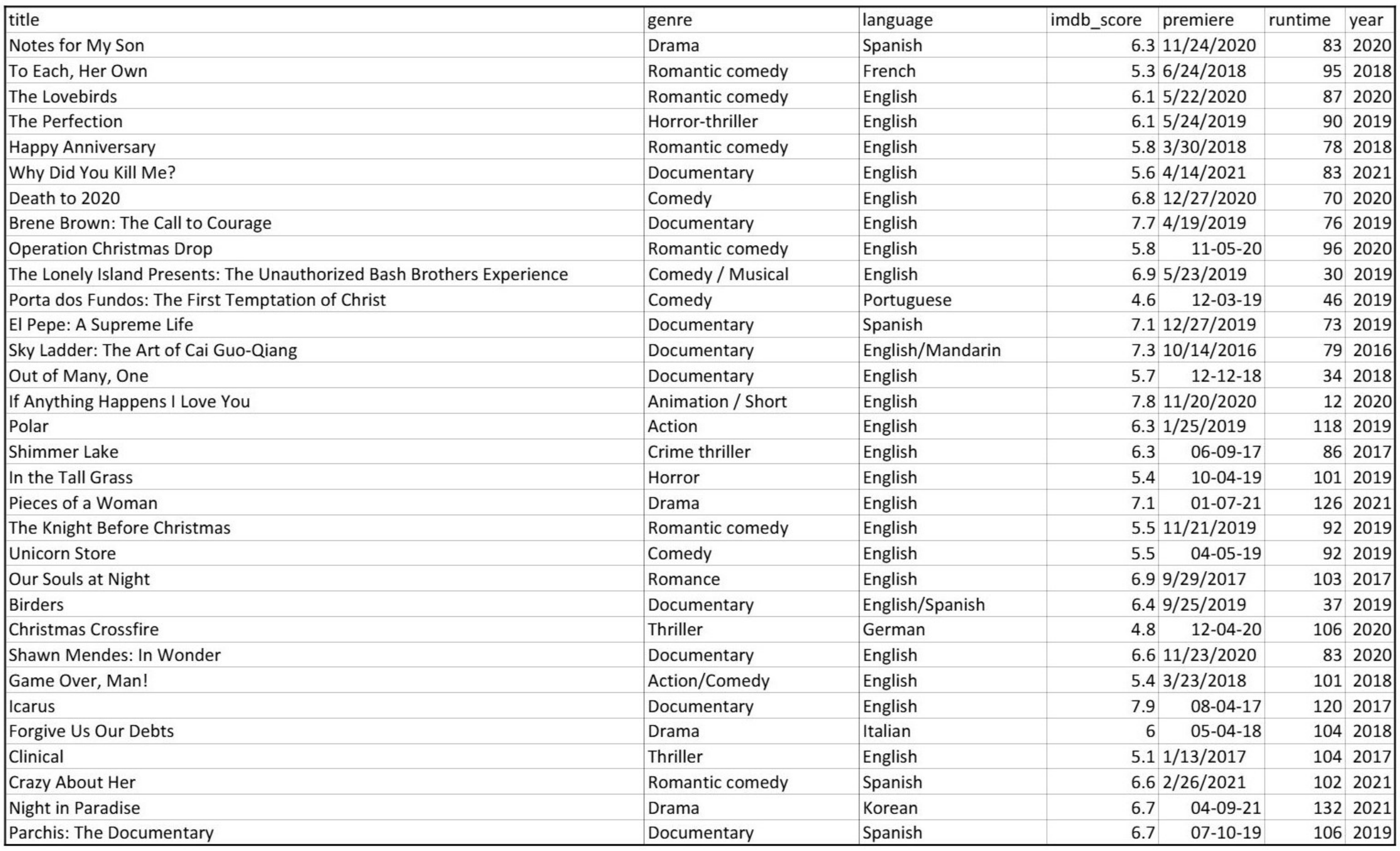
5.2. Content Distributor to Content Creator

Netflix utilizes the Open Connect Content Delivery Network as a Distributor taking care of the logistics for the delivery to different platforms and various devices. This ensures that content is available, anytime and anywhere, to all users. reducing the amount of latency and increases the streaming quality, provided that video content is cached closer to the users. To voice or not to voice, that is the question of which quality of a stream is paramount for content delivery in the digital sphere: being one of the first companies to use Cloud Services in the sphere that deals with big amounts of data and traffic, namely video streaming, which utilizes AWS. successful case of cloud computing scalability is the basis of next content distribution solutions, drawing benefits in relative ability to scale the resources has proven to be one of the most important service characteristics that stands at the heart of today's streaming landscape. Many cases distributor is responsible for properly monetizing the content through marketing/advertising, pay-per-views (PPV) and subscriptions packages with peaks. Netflix also negotiate and manages licensing and distributing rights in different areas and platform for the content.

from obtaining content from being a CD/DVD, Blue-Ray selling company and drastically changing the creative and commerce process of creating. acquiring content through partnerships with creators, production team, studios, and companies. negotiating agreements of sharing revenue and management rights on how content is made available. these changes are influenced with big data analytics and change the traditional industry of television and movie production. Moreover, in the past few years, Netflix has never been stingy with spending a huge amount on original productions, producing numerous series and films for the public. While this kindled focus continues to pay off in the form of Netflix additions to its content library, it also expanded the creation of original content in other streaming services in an effort to stand out from the competition. also, changed how companies think about the UX design and the features that should be included in a product. building a global outlook into everything like equity, diversity, inclusion and foster a culture of empathy, curiosity and courage. they develop faster and better story and to share to all the members around the world.

Section B

6. Datasets



title	genre	language	imdb_score	premiere	runtime	year
Notes for My Son	Drama	Spanish	6.3	11/24/2020	83	2020
To Each, Her Own	Romantic comedy	French	5.3	6/24/2018	95	2018
The Lovebirds	Romantic comedy	English	6.1	5/22/2020	87	2020
The Perfection	Horror-thriller	English	6.1	5/24/2019	90	2019
Happy Anniversary	Romantic comedy	English	5.8	3/30/2018	78	2018
Why Did You Kill Me?	Documentary	English	5.6	4/14/2021	83	2021
Death to 2020	Comedy	English	6.8	12/27/2020	70	2020
Brene Brown: The Call to Courage	Documentary	English	7.7	4/19/2019	76	2019
Operation Christmas Drop	Romantic comedy	English	5.8	11-05-20	96	2020
The Lonely Island Presents: The Unauthorized Bash Brothers Experience	Comedy / Musical	English	6.9	5/23/2019	30	2019
Porta dos Fundos: The First Temptation of Christ	Comedy	Portuguese	4.6	12-03-19	46	2019
El Pepe: A Supreme Life	Documentary	Spanish	7.1	12/27/2019	73	2019
Sky Ladder: The Art of Cai Guo-Qiang	Documentary	English/Mandarin	7.3	10/14/2016	79	2016
Out of Many, One	Documentary	English	5.7	12-12-18	34	2018
If Anything Happens I Love You	Animation / Short	English	7.8	11/20/2020	12	2020
Polar	Action	English	6.3	1/25/2019	118	2019
Shimmer Lake	Crime thriller	English	6.3	06-09-17	86	2017
In the Tall Grass	Horror	English	5.4	10-04-19	101	2019
Pieces of a Woman	Drama	English	7.1	01-07-21	126	2021
The Knight Before Christmas	Romantic comedy	English	5.5	11/21/2019	92	2019
Unicorn Store	Comedy	English	5.5	04-05-19	92	2019
Our Souls at Night	Romance	English	6.9	9/29/2017	103	2017
Birders	Documentary	English/Spanish	6.4	9/25/2019	37	2019
Christmas Crossfire	Thriller	German	4.8	12-04-20	106	2020
Shawn Mendes: In Wonder	Documentary	English	6.6	11/23/2020	83	2020
Game Over, Man!	Action/Comedy	English	5.4	3/23/2018	101	2018
Icarus	Documentary	English	7.9	08-04-17	120	2017
Forgive Us Our Debts	Drama	Italian	6	05-04-18	104	2018
Clinical	Thriller	English	5.1	1/13/2017	104	2017
Crazy About Her	Romantic comedy	Spanish	6.6	2/26/2021	102	2021
Night in Paradise	Drama	Korean	6.7	04-09-21	132	2021
Parchis: The Documentary	Documentary	Spanish	6.7	07-10-19	106	2019

Figure 2. Netflix Dataset

the chosen dataset, are widely available on various open-source platforms. Kaggle, GitHub Data Repositories, are very popular as they provide rich variety of meticulous, comprehensive formats. the CSV format is very easy to parse, use, analyze with many tools. we can find them by all rich community of data enthusiasts across the world and researchers who support and make it easy by using single point, keywords, and also their additional scripts to get a head start, with well documented academic research reports, metadata, aggregated

6.1. Rationale

analyzing the datset with provided data about trends, genres, langauages, etc can have meaningful insights with historical significance. these can enhance the algorithms, which ensure the user are presented with content they would love watching. this allow sound meetings for content strategy (prefered length) and one up over competitors and infer appeal and quality of contents improving marketing. local socio-cultural representaion on global platform scale.

6.1.1. Scalability

Hadoops ability to distribute data to clusters that ensure that large datasetsa are scalable to perabytes and can accommodate any increase for the future. HBases storage model of vertical columns allow for efficient handling for spares structures.which is easy for machine to work on as human have it easy with horizontal rows tables. Hive

Warehousing perform SQL-like queries, Pig Data Flow high level scripting MR parallel paradigm jobs across multiple nodes.

6.1.2. Ease of Integration

Hadoop ecosystem are designed for smooth integration and work together seamlessly compatibilty. Ingestion and Retrieval for large dataset is made easy with HBase. using familiar SQL like quires are accessable with Hive. processing and analyzing complex data transformation and workflows is simplified with Pig. customize writing of data handling, processing and analysis task is possible with MapReduce Framework

6.1.3. Compatibility with Current Systems

due to its open source nature, the hadoop ecosystem is backed up with large community support and with no licensing issues. it is very modular and is able to fit into various technological stack archietecture without nay major overhaul. models are easily accomodated with HBase and Hive for Netflix dataset structures. it is crossplat from meaning it compatibile on many different infrastructures and multiple operating systems.

6.2. Brief Introduction

The Netflix dataset for analysis provides comprehensive insights into the diverse content available on the Netflix platform. This dataset includes detailed information on various titles, covering key attributes such as genre, premiere date, runtime, IMDb score, language, and release year. Analyzing this data can help identify trends, improve recommendation systems, and enhance understanding of Netflix's content strategy.

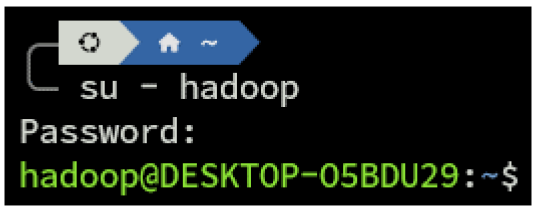
6.3. Parameters in the Table.

Parameter	Description	Data Types
Title	The title of the movie or TV show.	String
Genre	The category or type of content, indicating its theme or style.	String
Premiere	The date when the movie or TV show was first released or premiered.	Date
Runtime	The duration of the movie or TV show in minutes.	Int
IMDb Score	The rating of the movie or TV show on the IMDb platform, representing its overall quality as rated by users.	Float
Language	The language in which the movie or TV show is primarily spoken or produced.	Sting
Year	The year when the movie or TV show was released or premiered.	Int

7. Operations on the chosen dataset:

7.1. Hadoop

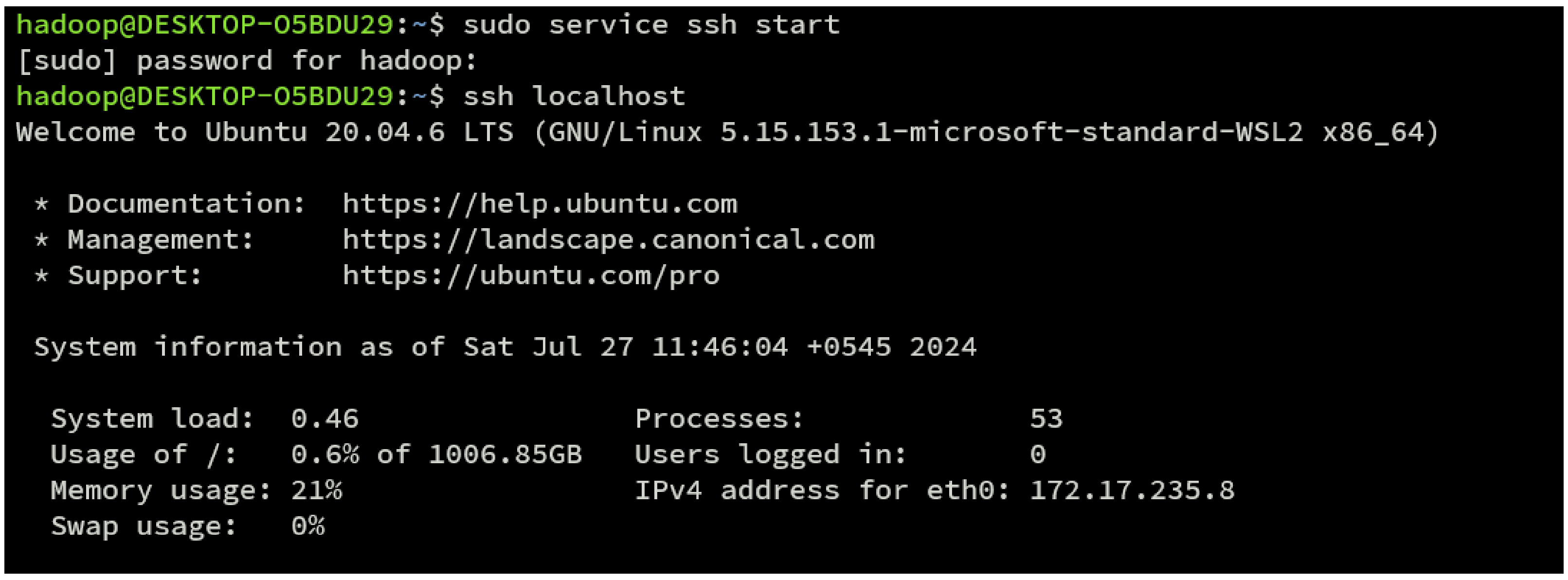
login into hadoop user



```
~
╰ su - hadoop
Password:
hadoop@DESKTOP-05BDU29:~$
```

Figure 3. Hadoop User Login

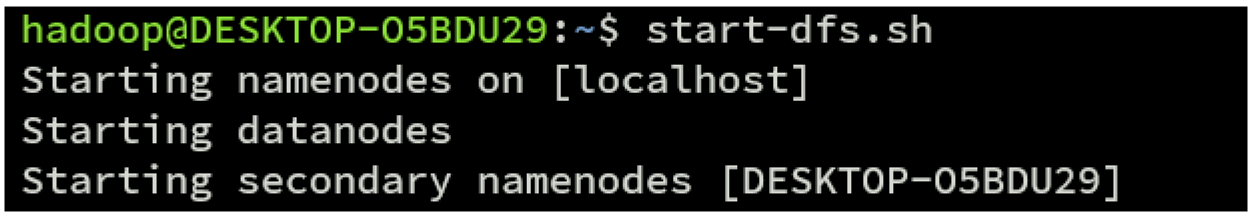
start ssh service to localhost



```
hadoop@DESKTOP-05BDU29:~$ sudo service ssh start
[sudo] password for hadoop:
hadoop@DESKTOP-05BDU29:~$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.153.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Sat Jul 27 11:46:04 +0545 2024

  System load:  0.46               Processes:             53
  Usage of /:   0.6% of 1006.85GB  Users logged in:       0
  Memory usage: 21%                IPv4 address for eth0: 172.17.235.8
  Swap usage:   0%
```

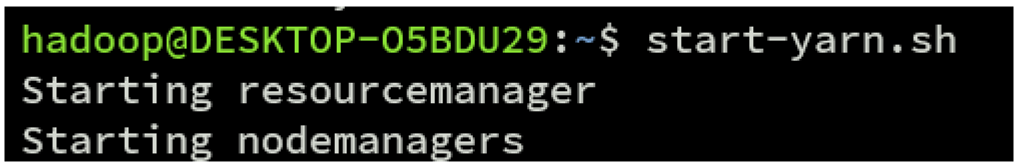
Figure 4. ssh localhost

start dfs, yarn, hbase, thrift, zookeeper server



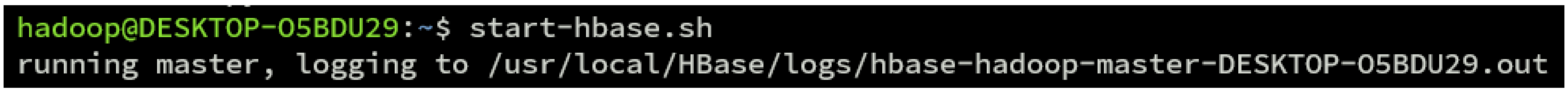
```
hadoop@DESKTOP-05BDU29:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [DESKTOP-05BDU29]
```

Figure 5. start-dfs



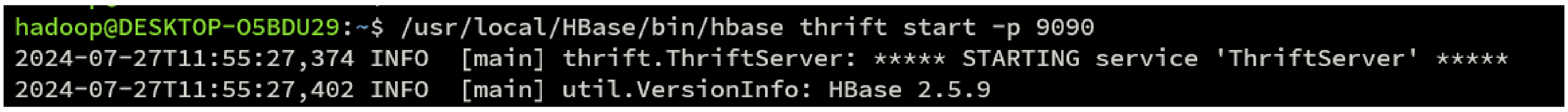
```
hadoop@DESKTOP-05BDU29:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

Figure 6. start-yarn



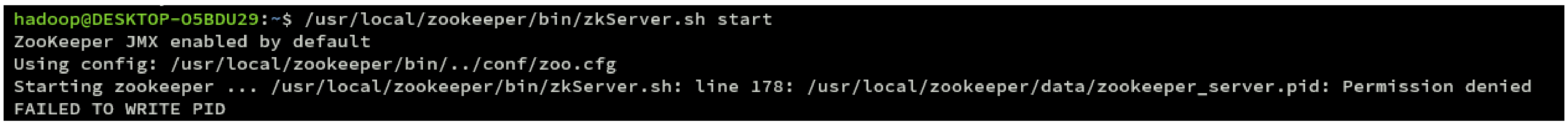
```
hadoop@DESKTOP-05BDU29:~$ start-hbase.sh
running master, logging to /usr/local/HBase/logs/hbase-hadoop-master-DESKTOP-05BDU29.out
```

Figure 7. start-hbase



```
hadoop@DESKTOP-05BDU29:~$ /usr/local/HBase/bin/hbase thrift start -p 9090
2024-07-27T11:55:27,374 INFO  [main] thrift.ThriftServer: ***** STARTING service 'ThriftServer' *****
2024-07-27T11:55:27,402 INFO  [main] util.VersionInfo: HBase 2.5.9
```

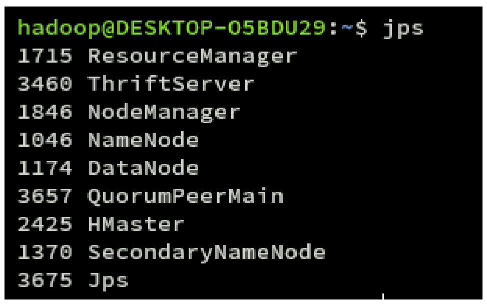
Figure 8. start-thrift



```
hadoop@DESKTOP-05BDU29:~$ /usr/local/zookeeper/bin/zkServer.sh start
ZooKeeper JMX enabled by default
Using config: /usr/local/zookeeper/bin/../conf/zoo.cfg
Starting zookeeper ... /usr/local/zookeeper/bin/zkServer.sh: line 178: /usr/local/zookeeper/data/zookeeper_server.pid: Permission denied
FAILED TO WRITE PID
```

Figure 9. start-zookeeper

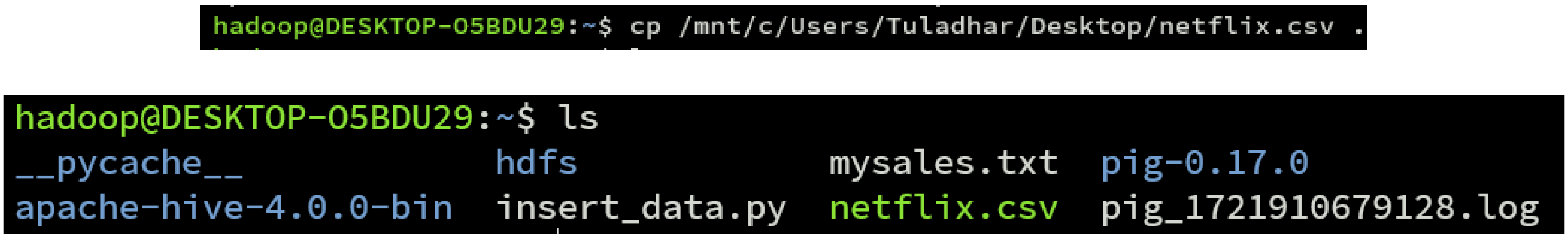
check jps



```
hadoop@DESKTOP-05BDU29:~$ jps
1715 ResourceManager
3460 ThriftServer
1846 NodeManager
1046 NameNode
1174 DataNode
3657 QuorumPeerMain
2425 HMaster
1370 SecondaryNameNode
3675 Jps
```

Figure 10. jps

copy dataset across windows and linux

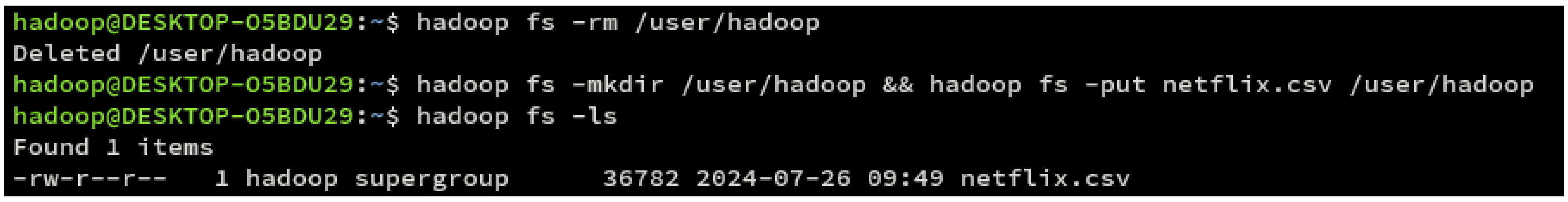


```
hadoop@DESKTOP-05BDU29:~$ cp /mnt/c/Users/Tuladhar/Desktop/netflix.csv .
```

```
hadoop@DESKTOP-05BDU29:~$ ls
__pycache__            hdfs            mysales.txt  pig-0.17.0
apache-hive-4.0.0-bin  insert_data.py  netflix.csv  pig_1721910679128.log
```

Figure 11. copy dataset from windows to linux

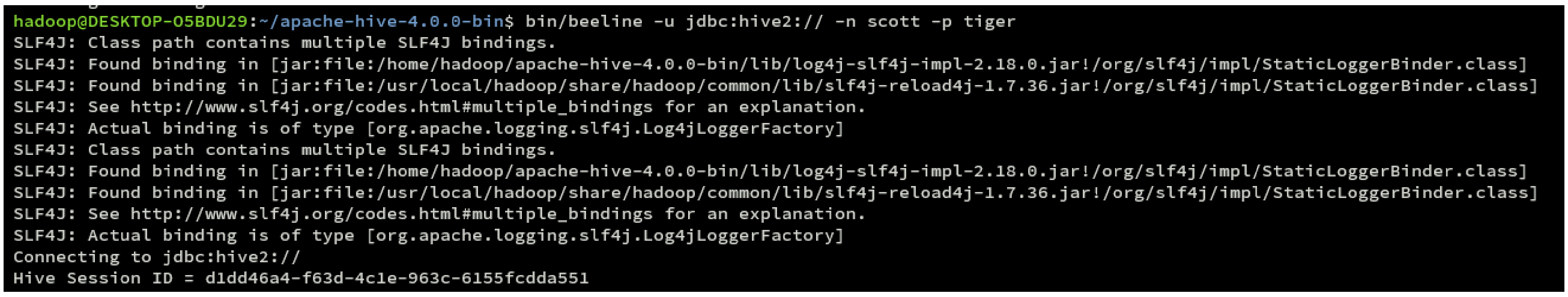
create remove hadoop diretory put file in hadoop



```
hadoop@DESKTOP-05BDU29:~$ hadoop fs -rm /user/hadoop
Deleted /user/hadoop
hadoop@DESKTOP-05BDU29:~$ hadoop fs -mkdir /user/hadoop && hadoop fs -put netflix.csv /user/hadoop
hadoop@DESKTOP-05BDU29:~$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 hadoop supergroup      36782 2024-07-26 09:49 netflix.csv
```

Figure 12. remove hadoop directory make new and put netflix csv

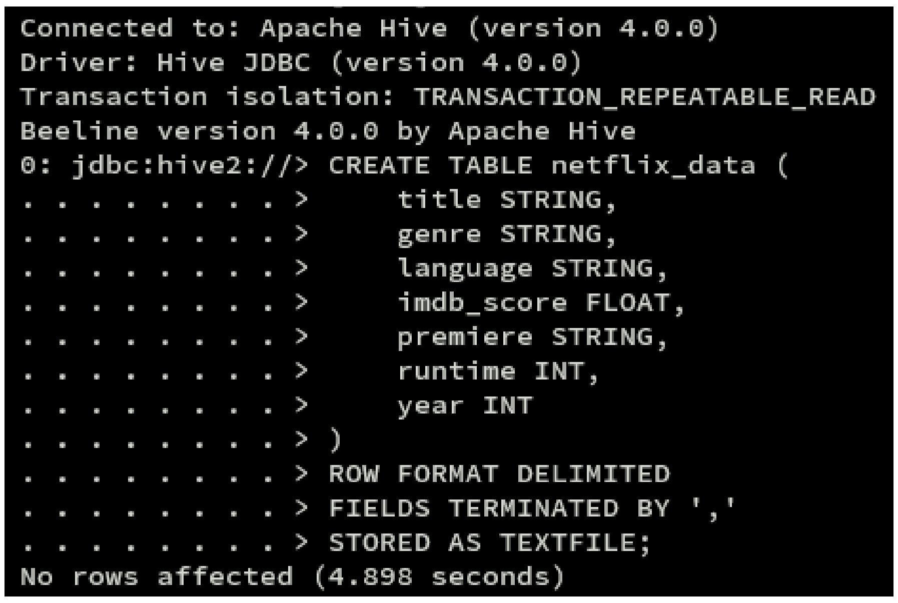
7.2. Hive



```
hadoop@DESKTOP-O5BDU29:~/apache-hive-4.0.0-bin$ bin/beeline -u jdbc:hive2:// -n scott -p tiger
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://
Hive Session ID = d1dd46a4-f63d-4c1e-963c-6155fcdda551
```

Figure 13. hive shell

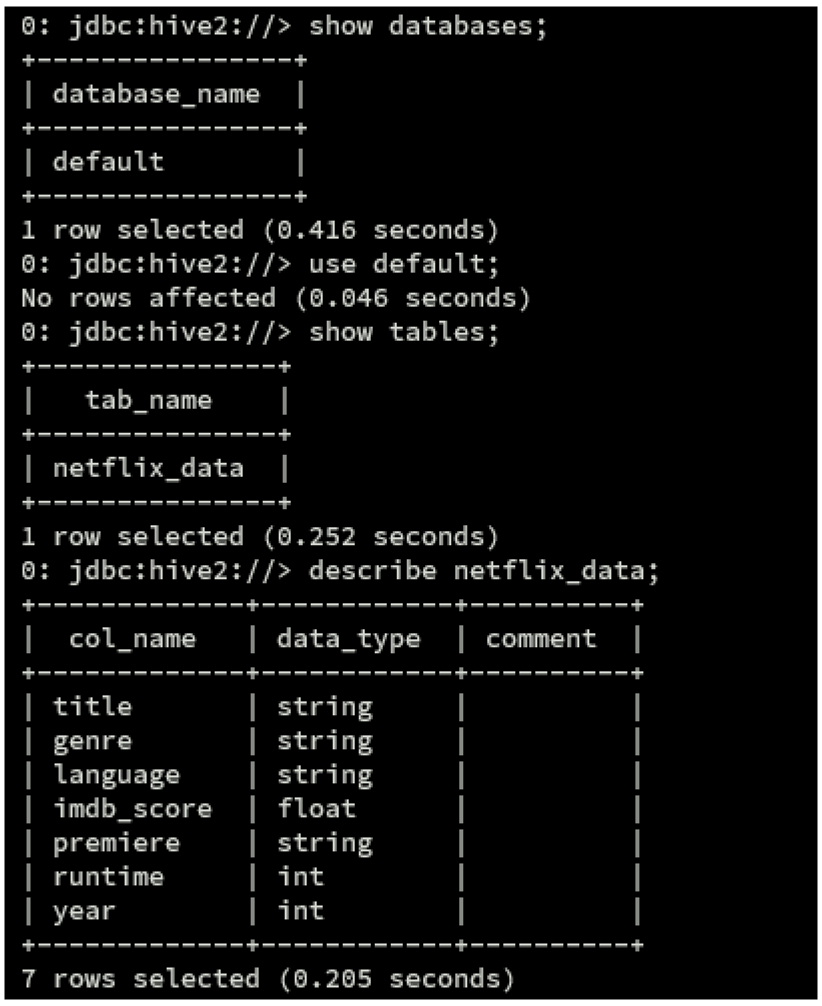
creation of table



```
Connected to: Apache Hive (version 4.0.0)
Driver: Hive JDBC (version 4.0.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 4.0.0 by Apache Hive
0: jdbc:hive2://> CREATE TABLE netflix_data (
. . . . . . . . >     title STRING,
. . . . . . . . >     genre STRING,
. . . . . . . . >     language STRING,
. . . . . . . . >     imdb_score FLOAT,
. . . . . . . . >     premiere STRING,
. . . . . . . . >     runtime INT,
. . . . . . . . >     year INT
. . . . . . . . > )
. . . . . . . . > ROW FORMAT DELIMITED
. . . . . . . . > FIELDS TERMINATED BY ','
. . . . . . . . > STORED AS TEXTFILE;
No rows affected (4.898 seconds)
```

Figure 14. make hive table

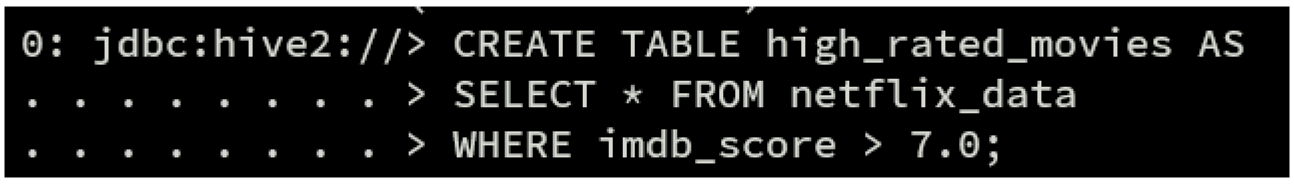
Show database, tables, describe Netflix data



```
0: jdbc:hive2://> show databases;
+----------------+
| database_name  |
+----------------+
| default        |
+----------------+
1 row selected (0.416 seconds)
0: jdbc:hive2://> use default;
No rows affected (0.046 seconds)
0: jdbc:hive2://> show tables;
+---------------+
|   tab_name    |
+---------------+
| netflix_data  |
+---------------+
1 row selected (0.252 seconds)
0: jdbc:hive2://> describe netflix_data;
+-------------+------------+----------+
|  col_name   | data_type  | comment  |
+-------------+------------+----------+
| title       | string     |          |
| genre       | string     |          |
| language    | string     |          |
| imdb_score  | float      |          |
| premiere    | string     |          |
| runtime     | int        |          |
| year        | int        |          |
+-------------+------------+----------+
7 rows selected (0.205 seconds)
```

Figure 15. show existing database, table, and describe

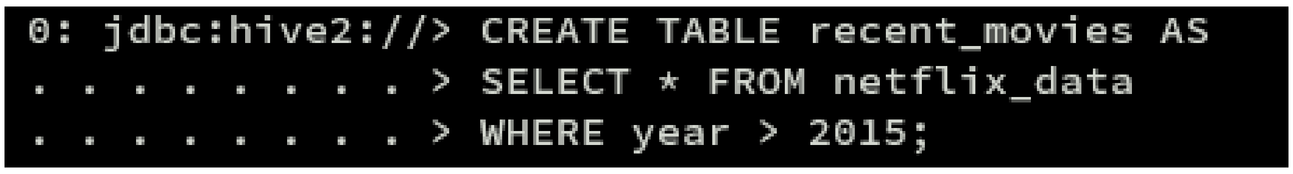
Query to filter movies with IMDB score greater than 7



```
0: jdbc:hive2://> CREATE TABLE high_rated_movies AS
. . . . . . . . > SELECT * FROM netflix_data
. . . . . . . . > WHERE imdb_score > 7.0;
```

Figure 16. create table show movies with score more than 7

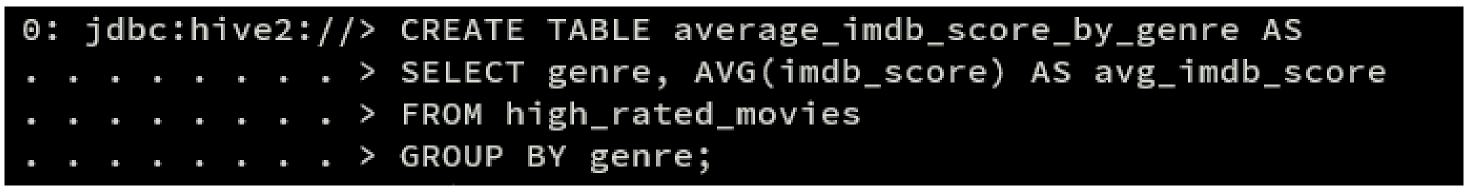
Create a table for movies released after 2015



```
0: jdbc:hive2://> CREATE TABLE recent_movies AS
. . . . . . . . > SELECT * FROM netflix_data
. . . . . . . . > WHERE year > 2015;
```

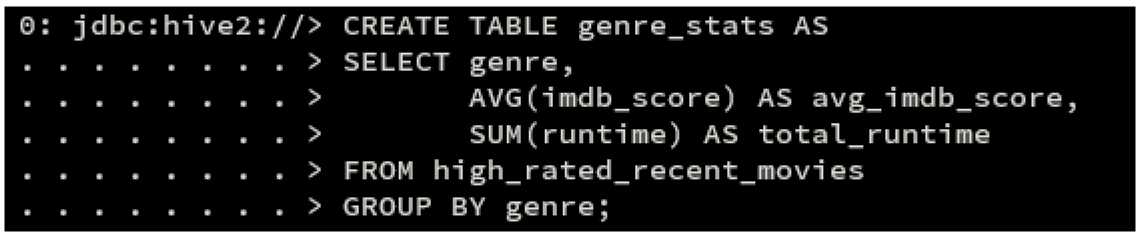
Figure 17. table for movies released after 2015

Query to group by genre and calculate average IMDB score and total runtime for each genre



```
0: jdbc:hive2://> CREATE TABLE average_imdb_score_by_genre AS
. . . . . . . . > SELECT genre, AVG(imdb_score) AS avg_imdb_score
. . . . . . . . > FROM high_rated_movies
. . . . . . . . > GROUP BY genre;
```

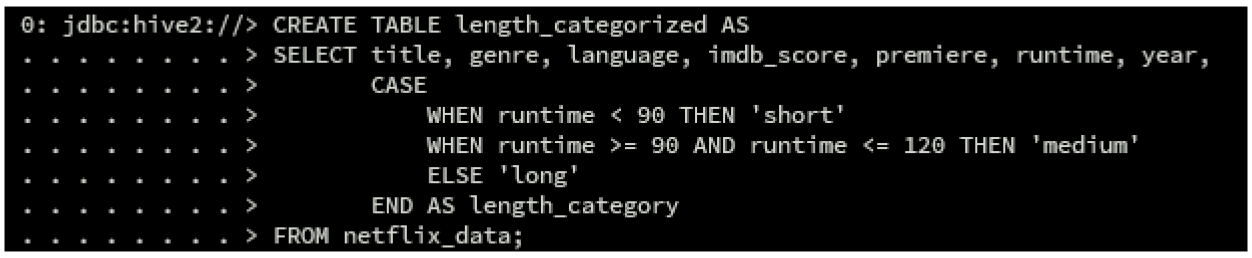
Figure 18.table for average imbd score by genre



```
0: jdbc:hive2://> CREATE TABLE genre_stats AS
. . . . . . . . > SELECT genre,
. . . . . . . . >        AVG(imdb_score) AS avg_imdb_score,
. . . . . . . . >        SUM(runtime) AS total_runtime
. . . . . . . . > FROM high_rated_recent_movies
. . . . . . . . > GROUP BY genre;
```

Figure 19. average imdb score and runtime

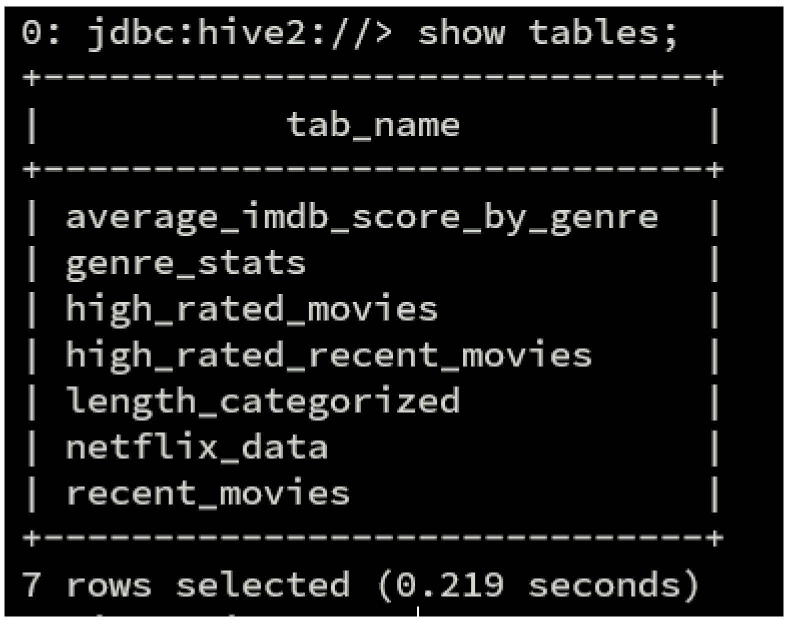
Create a derived column for length category (short, medium, long)



```
0: jdbc:hive2://> CREATE TABLE length_categorized AS
. . . . . . . . > SELECT title, genre, language, imdb_score, premiere, runtime, year,
. . . . . . . . >        CASE
. . . . . . . . >            WHEN runtime < 90 THEN 'short'
. . . . . . . . >            WHEN runtime >= 90 AND runtime <= 120 THEN 'medium'
. . . . . . . . >            ELSE 'long'
. . . . . . . . >        END AS length_category
. . . . . . . . > FROM netflix_data;
```

Figure 20. derived column for lengths (S,M,L)

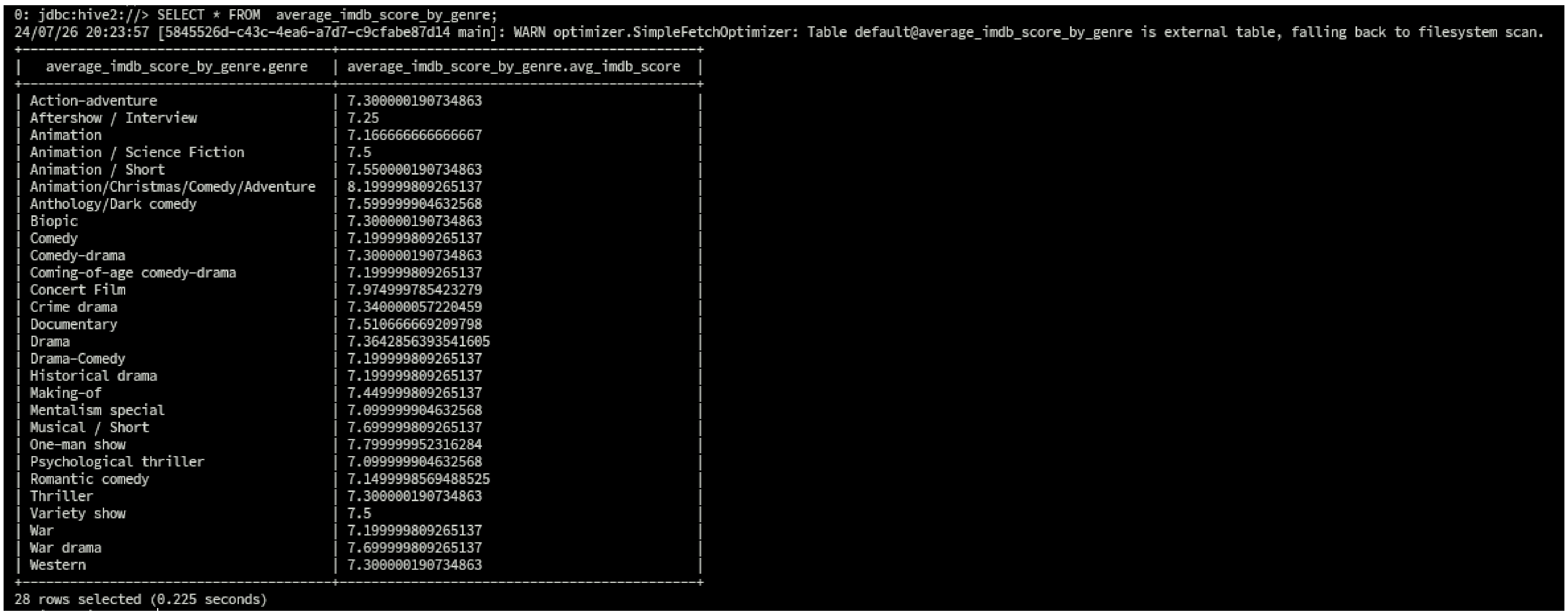
show tables result after



```
0: jdbc:hive2://> show tables;
+-----------------------------+
|          tab_name           |
+-----------------------------+
| average_imdb_score_by_genre |
| genre_stats                 |
| high_rated_movies           |
| high_rated_recent_movies    |
| length_categorized          |
| netflix_data                |
| recent_movies               |
+-----------------------------+
7 rows selected (0.219 seconds)
```

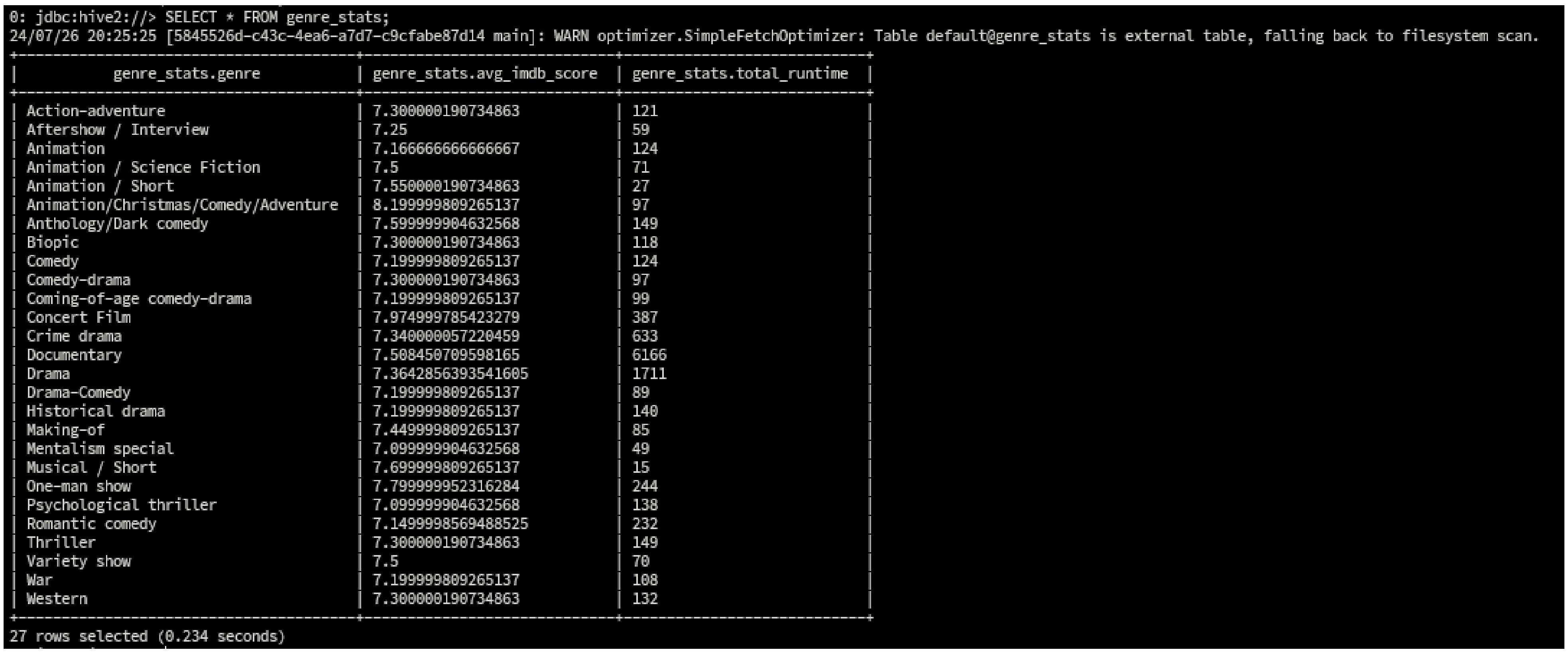
Figure 21. show made tables

Select from the result tables SELECT * FROM



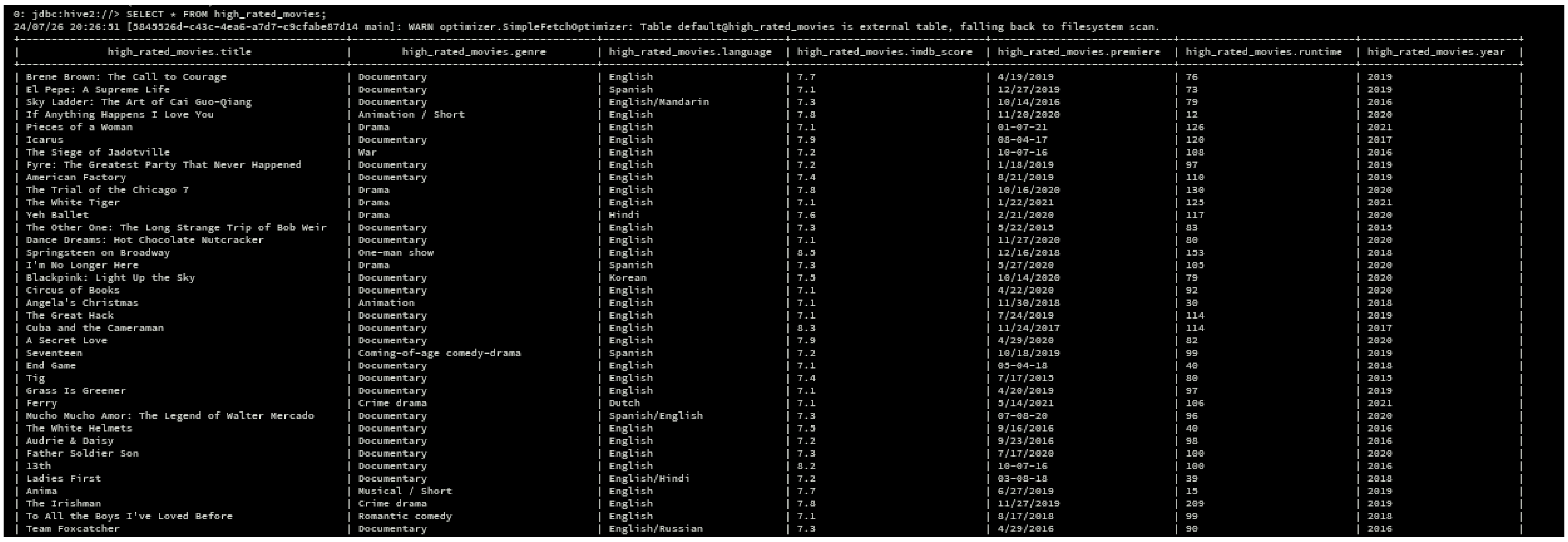
```
0: jdbc:hive2://> SELECT * FROM  average_imdb_score_by_genre;
24/07/26 20:23:57 [5845526d-c43c-4ea6-a7d7-c9cfabe87d14 main]: WARN optimizer.SimpleFetchOptimizer: Table default@average_imdb_score_by_genre is external table, falling back to filesystem scan.
+-------------------------------------+-------------------------------------------+
|   average_imdb_score_by_genre.genre  | average_imdb_score_by_genre.avg_imdb_score  |
+-------------------------------------+-------------------------------------------+
| Action-adventure                    | 7.300000190734863                         |
| Aftershow / Interview               | 7.25                                      |
| Animation                           | 7.166666666666667                         |
| Animation / Science Fiction         | 7.5                                       |
| Animation / Short                   | 7.550000190734863                         |
| Animation/Christmas/Comedy/Adventure | 8.199999809265137                         |
| Anthology/Dark comedy               | 7.599999904632568                         |
| Biopic                              | 7.300000190734863                         |
| Comedy                              | 7.199999809265137                         |
| Comedy-drama                        | 7.300000190734863                         |
| Coming-of-age comedy-drama          | 7.199999809265137                         |
| Concert Film                        | 7.974999785423279                         |
| Crime drama                         | 7.340000057220459                         |
| Documentary                         | 7.510666669209798                         |
| Drama                               | 7.3642856393541605                        |
| Drama-Comedy                        | 7.199999809265137                         |
| Historical drama                    | 7.199999809265137                         |
| Making-of                           | 7.449999809265137                         |
| Mentalism special                   | 7.099999904632568                         |
| Musical / Short                     | 7.699999809265137                         |
| One-man show                        | 7.799999952316284                         |
| Psychological thriller              | 7.099999904632568                         |
| Romantic comedy                     | 7.1499998569488525                        |
| Thriller                            | 7.300000190734863                         |
| Variety show                        | 7.5                                       |
| War                                 | 7.199999809265137                         |
| War drama                           | 7.699999809265137                         |
| Western                             | 7.300000190734863                         |
+-------------------------------------+-------------------------------------------+
28 rows selected (0.225 seconds)
```

Figure 22. average imdb score by genre table



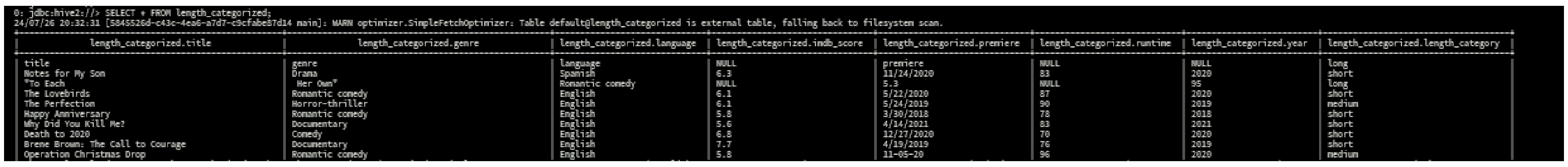
```
0: jdbc:hive2://> SELECT * FROM genre_stats;
24/07/26 20:25:25 [5845526d-c43c-4ea6-a7d7-c9cfabe87d14 main]: WARN optimizer.SimpleFetchOptimizer: Table default@genre_stats is external table, falling back to filesystem scan.
+---------------------------------------+-----------------------------+---------------------------+
|          genre_stats.genre            | genre_stats.avg_imdb_score  | genre_stats.total_runtime |
+---------------------------------------+-----------------------------+---------------------------+
| Action-adventure                      | 7.300000190734863           | 121                       |
| Aftershow / Interview                 | 7.25                        | 59                        |
| Animation                             | 7.166666666666667           | 124                       |
| Animation / Science Fiction           | 7.5                         | 71                        |
| Animation / Short                     | 7.550000190734863           | 27                        |
| Animation/Christmas/Comedy/Adventure  | 8.199999809265137           | 97                        |
| Anthology/Dark comedy                 | 7.599999904632568           | 149                       |
| Biopic                                | 7.300000190734863           | 118                       |
| Comedy                                | 7.199999809265137           | 124                       |
| Comedy-drama                          | 7.300000190734863           | 97                        |
| Coming-of-age comedy-drama            | 7.199999809265137           | 99                        |
| Concert Film                          | 7.974999785423279           | 387                       |
| Crime drama                           | 7.340000057220459           | 633                       |
| Documentary                           | 7.508450709598165           | 6166                      |
| Drama                                 | 7.3642856393541605          | 1711                      |
| Drama-Comedy                          | 7.199999809265137           | 89                        |
| Historical drama                      | 7.199999809265137           | 140                       |
| Making-of                             | 7.449999809265137           | 85                        |
| Mentalism special                     | 7.099999904632568           | 49                        |
| Musical / Short                       | 7.699999809265137           | 15                        |
| One-man show                          | 7.799999952316284           | 244                       |
| Psychological thriller                | 7.099999904632568           | 138                       |
| Romantic comedy                       | 7.1499998569488525          | 232                       |
| Thriller                              | 7.300000190734863           | 149                       |
| Variety show                          | 7.5                         | 70                        |
| War                                   | 7.199999809265137           | 108                       |
| Western                               | 7.300000190734863           | 132                       |
+---------------------------------------+-----------------------------+---------------------------+
27 rows selected (0.234 seconds)
```

Figure 23. genre stats table



```
0: jdbc:hive2://> SELECT * FROM high_rated_movies;
24/07/26 20:26:51 [5845526d-c43c-4ea6-a7d7-c9cfabe87d14 main]: WARN optimizer.SimpleFetchOptimizer: Table default@high_rated_movies is external table, falling back to filesystem scan.
+---------------------------------------------------+-------------------------------------+-----------------------------+--------------------------------+-------------------------------+------------------------------+---------------------------+
|              high_rated_movies.title              |        high_rated_movies.genre      | high_rated_movies.language  | high_rated_movies.imdb_score   | high_rated_movies.premiere    | high_rated_movies.runtime    | high_rated_movies.year    |
+---------------------------------------------------+-------------------------------------+-----------------------------+--------------------------------+-------------------------------+------------------------------+---------------------------+
| Brene Brown: The Call to Courage                  | Documentary                         | English                     | 7.7                            | 4/19/2019                     | 76                           | 2019                      |
| El Pepe: A Supreme Life                           | Documentary                         | Spanish                     | 7.1                            | 12/27/2019                    | 73                           | 2019                      |
| Sky Ladder: The Art of Cai Guo-Qiang              | Documentary                         | English/Mandarin            | 7.3                            | 10/14/2016                    | 79                           | 2016                      |
| If Anything Happens I Love You                    | Animation / Short                   | English                     | 7.8                            | 11/20/2020                    | 12                           | 2020                      |
| Pieces of a Woman                                 | Drama                               | English                     | 7.1                            | 01-07-21                      | 126                          | 2021                      |
| Icarus                                            | Documentary                         | English                     | 7.9                            | 08-04-17                      | 120                          | 2017                      |
| The Siege of Jadotville                           | War                                 | English                     | 7.2                            | 10-07-16                      | 108                          | 2016                      |
| Fyre: The Greatest Party That Never Happened      | Documentary                         | English                     | 7.2                            | 1/18/2019                     | 97                           | 2019                      |
| American Factory                                  | Documentary                         | English                     | 7.4                            | 8/21/2019                     | 110                          | 2019                      |
| The Trial of the Chicago 7                        | Drama                               | English                     | 7.8                            | 10/16/2020                    | 130                          | 2020                      |
| The White Tiger                                   | Drama                               | English                     | 7.1                            | 1/22/2021                     | 125                          | 2021                      |
| Yeh Ballet                                        | Drama                               | Hindi                       | 7.6                            | 2/21/2020                     | 117                          | 2020                      |
| The Other One: The Long Strange Trip of Bob Weir  | Documentary                         | English                     | 7.3                            | 5/22/2015                     | 83                           | 2015                      |
| Dance Dreams: Hot Chocolate Nutcracker            | Documentary                         | English                     | 7.1                            | 11/27/2020                    | 80                           | 2020                      |
| Springsteen on Broadway                           | One-man show                        | English                     | 8.5                            | 12/16/2018                    | 153                          | 2018                      |
| I'm No Longer Here                                | Drama                               | Spanish                     | 7.3                            | 5/27/2020                     | 105                          | 2020                      |
| Blackpink: Light Up the Sky                       | Documentary                         | Korean                      | 7.5                            | 10/14/2020                    | 79                           | 2020                      |
| Circus of Books                                   | Documentary                         | English                     | 7.1                            | 4/22/2020                     | 92                           | 2020                      |
| Angela's Christmas                                | Animation                           | English                     | 7.1                            | 11/30/2018                    | 30                           | 2018                      |
| The Great Hack                                    | Documentary                         | English                     | 7.1                            | 7/24/2019                     | 114                          | 2019                      |
| Cuba and the Cameraman                            | Documentary                         | English                     | 8.3                            | 11/24/2017                    | 114                          | 2017                      |
| A Secret Love                                     | Documentary                         | English                     | 7.9                            | 4/29/2020                     | 82                           | 2020                      |
| Seventeen                                         | Coming-of-age comedy-drama          | Spanish                     | 7.2                            | 10/18/2019                    | 99                           | 2019                      |
| End Game                                          | Documentary                         | English                     | 7.1                            | 05-04-18                      | 40                           | 2018                      |
| Tig                                               | Documentary                         | English                     | 7.4                            | 7/17/2015                     | 80                           | 2015                      |
| Grass Is Greener                                  | Documentary                         | English                     | 7.1                            | 4/20/2019                     | 97                           | 2019                      |
| Ferry                                             | Crime drama                         | Dutch                       | 7.1                            | 5/14/2021                     | 106                          | 2021                      |
| Mucho Mucho Amor: The Legend of Walter Mercado    | Documentary                         | Spanish/English             | 7.3                            | 07-08-20                      | 96                           | 2020                      |
| The White Helmets                                 | Documentary                         | English                     | 7.5                            | 9/16/2016                     | 40                           | 2016                      |
| Audrie & Daisy                                    | Documentary                         | English                     | 7.2                            | 9/23/2016                     | 98                           | 2016                      |
| Father Soldier Son                                | Documentary                         | English                     | 7.3                            | 7/17/2020                     | 100                          | 2020                      |
| 13th                                              | Documentary                         | English                     | 8.2                            | 10-07-16                      | 100                          | 2016                      |
| Ladies First                                      | Documentary                         | English/Hindi               | 7.2                            | 03-08-18                      | 39                           | 2018                      |
| Anima                                             | Musical / Short                     | English                     | 7.7                            | 6/27/2019                     | 15                           | 2019                      |
| The Irishman                                      | Crime drama                         | English                     | 7.8                            | 11/27/2019                    | 209                          | 2019                      |
| To All the Boys I've Loved Before                 | Romantic comedy                     | English                     | 7.1                            | 8/17/2018                     | 99                           | 2018                      |
| Team Foxcatcher                                   | Documentary                         | English/Russian             | 7.3                            | 4/29/2016                     | 90                           | 2016                      |
```

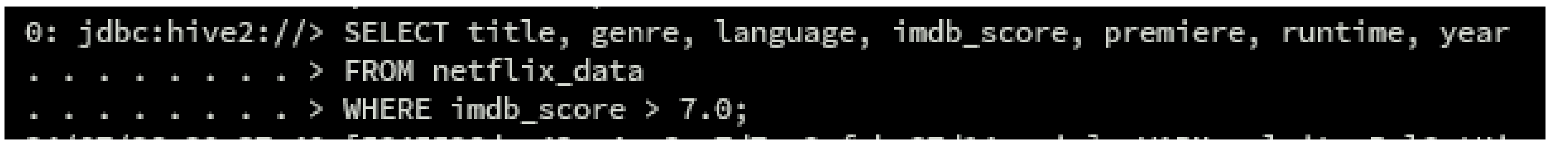
Figure 24. high rated movies table



```
0: jdbc:hive2://> SELECT * FROM length_categorized;
24/07/26 20:32:31 [5845526d-c43c-4ea6-a7d7-c9cfabe87d14 main]: WARN optimizer.SimpleFetchOptimizer: Table default@length_categorized is external table, falling back to filesystem scan.
+--------------------------------------+----------------------------------+-----------------------------+---------------------------------+--------------------------------+-------------------------------+----------------------------+-----------------------------------------+
|       length_categorized.title       |     length_categorized.genre     | length_categorized.language | length_categorized.imdb_score   | length_categorized.premiere    | length_categorized.runtime    | length_categorized.year    | length_categorized.length_category      |
+--------------------------------------+----------------------------------+-----------------------------+---------------------------------+--------------------------------+-------------------------------+----------------------------+-----------------------------------------+
| title                                | genre                            | language                    | NULL                            | premiere                       | NULL                          | NULL                       | long                                    |
  Notes for My Son                     | Drama                            | Spanish                     | 6.3                             | 11/24/2020                     | 83                            | 2020                       | short
  "To Each                             |  Her Own"                        | Romantic comedy             | NULL                            | 5.3                            | NULL                          | 95                         | long
  The Lovebirds                        | Romantic comedy                  | English                     | 6.1                             | 5/22/2020                      | 87                            | 2020                       | short
  The Perfection                       | Horror-thriller                  | English                     | 6.1                             | 5/24/2019                      | 90                            | 2019                       | medium
  Happy Anniversary                    | Romantic comedy                  | English                     | 5.8                             | 3/30/2018                      | 78                            | 2018                       | short
  Why Did You Kill Me?                 | Documentary                      | English                     | 5.6                             | 4/14/2021                      | 83                            | 2021                       | short
  Death to 2020                        | Comedy                           | English                     | 6.8                             | 12/27/2020                     | 70                            | 2020                       | short
  Brene Brown: The Call to Courage     | Documentary                      | English                     | 7.7                             | 4/19/2019                      | 76                            | 2019                       | short
  Operation Christmas Drop             | Romantic comedy                  | English                     | 5.8                             | 11-05-20                       | 96                            | 2020                       | medium
```

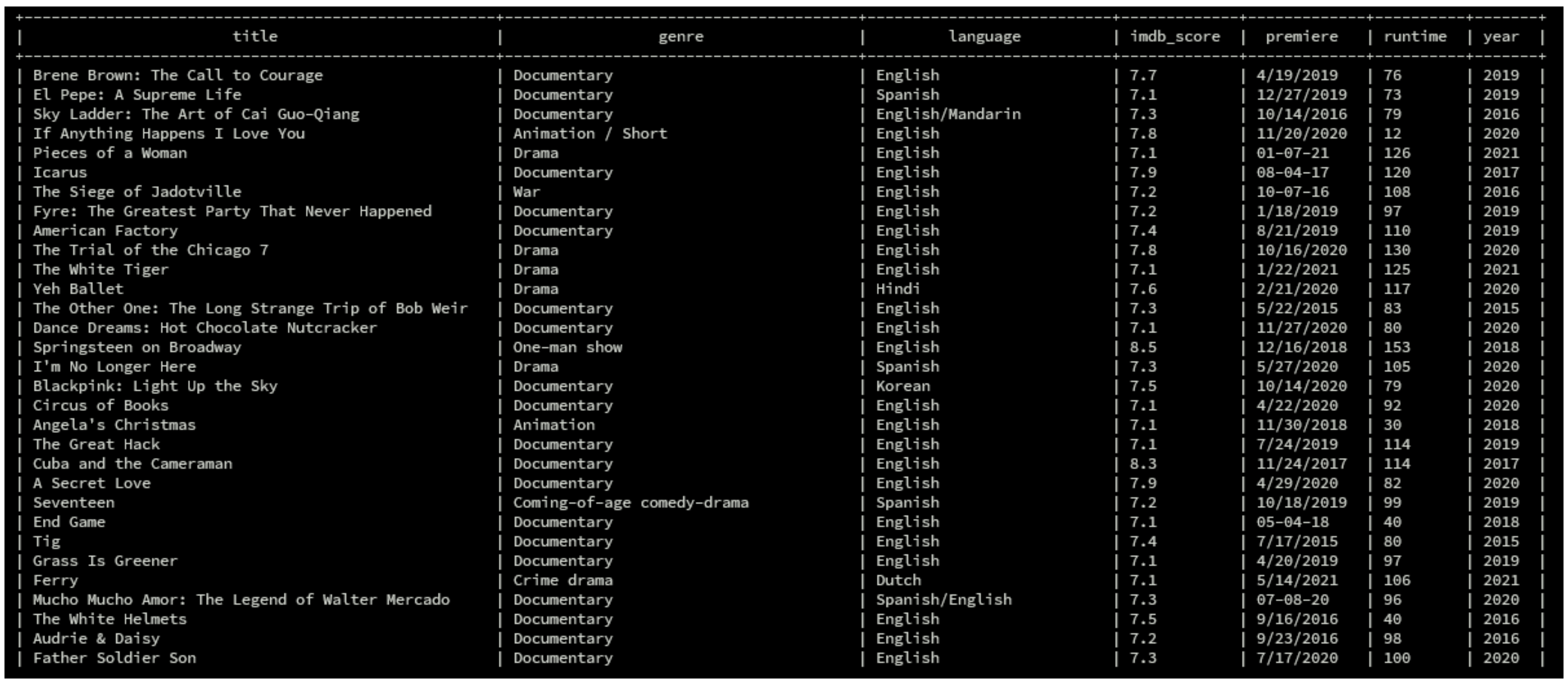
Figure 25. length categorized

flitering data with imdb score more than 7



```
0: jdbc:hive2://> SELECT title, genre, language, imdb_score, premiere, runtime, year
. . . . . . . . > FROM netflix_data
. . . . . . . . > WHERE imdb_score > 7.0;
```

Figure 26. filter data with score more than 7

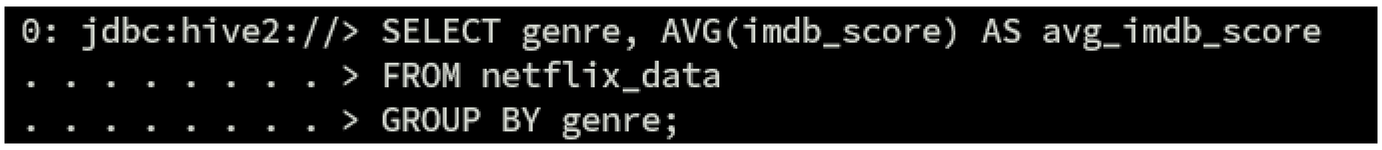


```
+--------------------------------------------------+------------------------------------+--------------------------+-------------+-------------+----------+-------+
|                      title                       |               genre                |         language         | imdb_score  |  premiere   | runtime  | year  |
+--------------------------------------------------+------------------------------------+--------------------------+-------------+-------------+----------+-------+
| Brene Brown: The Call to Courage                 | Documentary                        | English                  | 7.7         | 4/19/2019   | 76       | 2019  |
| El Pepe: A Supreme Life                          | Documentary                        | Spanish                  | 7.1         | 12/27/2019  | 73       | 2019  |
| Sky Ladder: The Art of Cai Guo-Qiang             | Documentary                        | English/Mandarin         | 7.3         | 10/14/2016  | 79       | 2016  |
| If Anything Happens I Love You                   | Animation / Short                  | English                  | 7.8         | 11/20/2020  | 12       | 2020  |
| Pieces of a Woman                                | Drama                              | English                  | 7.1         | 01-07-21    | 126      | 2021  |
| Icarus                                           | Documentary                        | English                  | 7.9         | 08-04-17    | 120      | 2017  |
| The Siege of Jadotville                          | War                                | English                  | 7.2         | 10-07-16    | 108      | 2016  |
| Fyre: The Greatest Party That Never Happened     | Documentary                        | English                  | 7.2         | 1/18/2019   | 97       | 2019  |
| American Factory                                 | Documentary                        | English                  | 7.4         | 8/21/2019   | 110      | 2019  |
| The Trial of the Chicago 7                       | Drama                              | English                  | 7.8         | 10/16/2020  | 130      | 2020  |
| The White Tiger                                  | Drama                              | English                  | 7.1         | 1/22/2021   | 125      | 2021  |
| Yeh Ballet                                       | Drama                              | Hindi                    | 7.6         | 2/21/2020   | 117      | 2020  |
| The Other One: The Long Strange Trip of Bob Weir | Documentary                        | English                  | 7.3         | 5/22/2015   | 83       | 2015  |
| Dance Dreams: Hot Chocolate Nutcracker           | Documentary                        | English                  | 7.1         | 11/27/2020  | 80       | 2020  |
| Springsteen on Broadway                          | One-man show                       | English                  | 8.5         | 12/16/2018  | 153      | 2018  |
| I'm No Longer Here                               | Drama                              | Spanish                  | 7.3         | 5/27/2020   | 105      | 2020  |
| Blackpink: Light Up the Sky                      | Documentary                        | Korean                   | 7.5         | 10/14/2020  | 79       | 2020  |
| Circus of Books                                  | Documentary                        | English                  | 7.1         | 4/22/2020   | 92       | 2020  |
| Angela's Christmas                               | Animation                          | English                  | 7.1         | 11/30/2018  | 30       | 2018  |
| The Great Hack                                   | Documentary                        | English                  | 7.1         | 7/24/2019   | 114      | 2019  |
| Cuba and the Cameraman                           | Documentary                        | English                  | 8.3         | 11/24/2017  | 114      | 2017  |
| A Secret Love                                    | Documentary                        | English                  | 7.9         | 4/29/2020   | 82       | 2020  |
| Seventeen                                        | Coming-of-age comedy-drama         | Spanish                  | 7.2         | 10/18/2019  | 99       | 2019  |
| End Game                                         | Documentary                        | English                  | 7.1         | 05-04-18    | 40       | 2018  |
| Tig                                              | Documentary                        | English                  | 7.4         | 7/17/2015   | 80       | 2015  |
| Grass Is Greener                                 | Documentary                        | English                  | 7.1         | 4/20/2019   | 97       | 2019  |
| Ferry                                            | Crime drama                        | Dutch                    | 7.1         | 5/14/2021   | 106      | 2021  |
| Mucho Mucho Amor: The Legend of Walter Mercado   | Documentary                        | Spanish/English          | 7.3         | 07-08-20    | 96       | 2020  |
| The White Helmets                                | Documentary                        | English                  | 7.5         | 9/16/2016   | 40       | 2016  |
| Audrie & Daisy                                   | Documentary                        | English                  | 7.2         | 9/23/2016   | 98       | 2016  |
| Father Soldier Son                               | Documentary                        | English                  | 7.3         | 7/17/2020   | 100      | 2020  |
```

Figure 27. output data with score more than 7

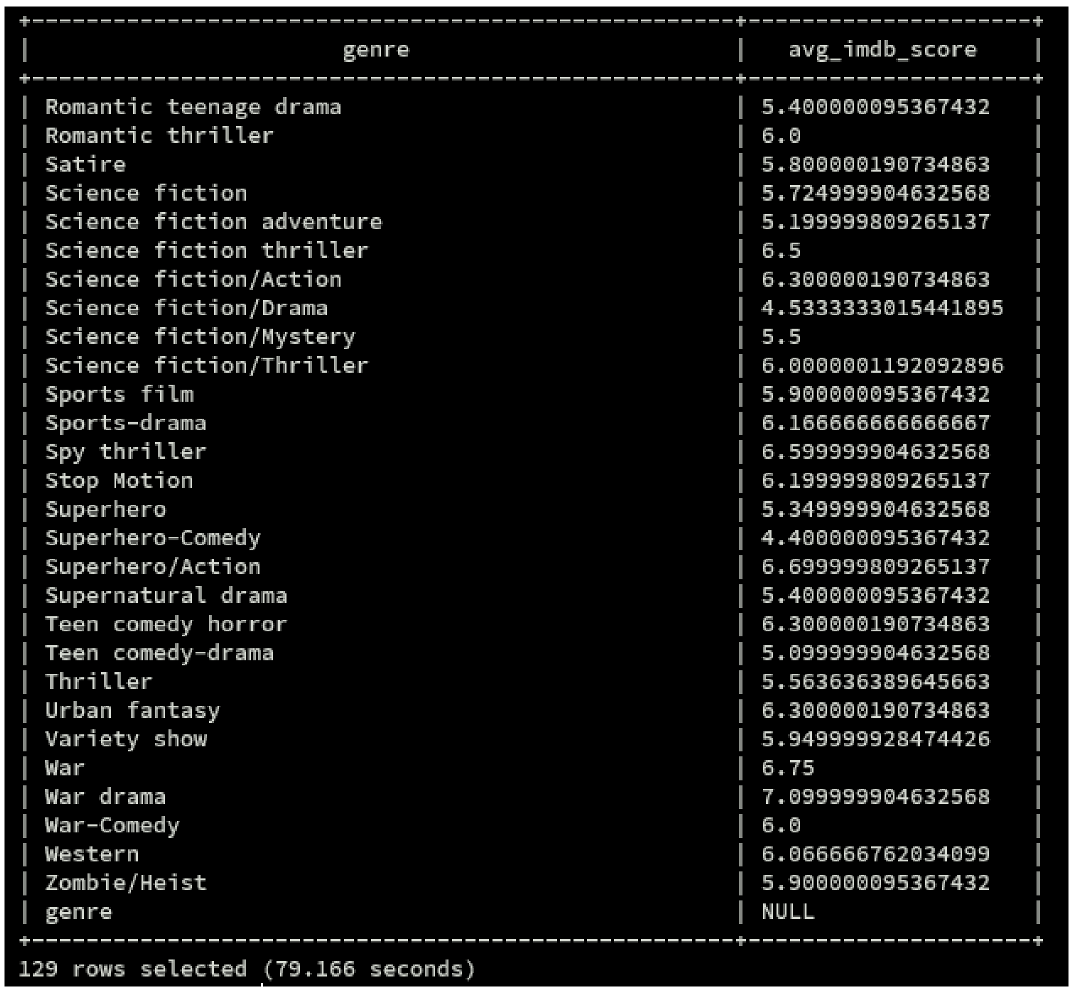
Aggregation Queries

Calculate the average IMDB score for each genre:



```
0: jdbc:hive2://> SELECT genre, AVG(imdb_score) AS avg_imdb_score
. . . . . . . . > FROM netflix_data
. . . . . . . . > GROUP BY genre;
```

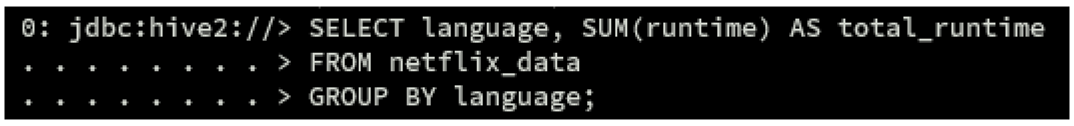
Figure 28. filter average imdb score for genre



```
+--------------------------------------------------+----------------------+
|                      genre                       |    avg_imdb_score    |
+--------------------------------------------------+----------------------+
| Romantic teenage drama                           | 5.400000095367432    |
| Romantic thriller                                | 6.0                  |
| Satire                                           | 5.800000190734863    |
| Science fiction                                  | 5.724999904632568    |
| Science fiction adventure                        | 5.199999809265137    |
| Science fiction thriller                         | 6.5                  |
| Science fiction/Action                           | 6.300000190734863    |
| Science fiction/Drama                            | 4.5333333015441895   |
| Science fiction/Mystery                          | 5.5                  |
| Science fiction/Thriller                         | 6.0000001192092896   |
| Sports film                                      | 5.900000095367432    |
| Sports-drama                                     | 6.166666666666667    |
| Spy thriller                                     | 6.599999904632568    |
| Stop Motion                                      | 6.199999809265137    |
| Superhero                                        | 5.349999904632568    |
| Superhero-Comedy                                 | 4.400000095367432    |
| Superhero/Action                                 | 6.699999809265137    |
| Supernatural drama                               | 5.400000095367432    |
| Teen comedy horror                               | 6.300000190734863    |
| Teen comedy-drama                                | 5.099999904632568    |
| Thriller                                         | 5.563636389645663    |
| Urban fantasy                                    | 6.300000190734863    |
| Variety show                                     | 5.949999928474426    |
| War                                              | 6.75                 |
| War drama                                        | 7.099999904632568    |
| War-Comedy                                       | 6.0                  |
| Western                                          | 6.066666762034099    |
| Zombie/Heist                                     | 5.900000095367432    |
| genre                                            | NULL                 |
+--------------------------------------------------+----------------------+
129 rows selected (79.166 seconds)
```

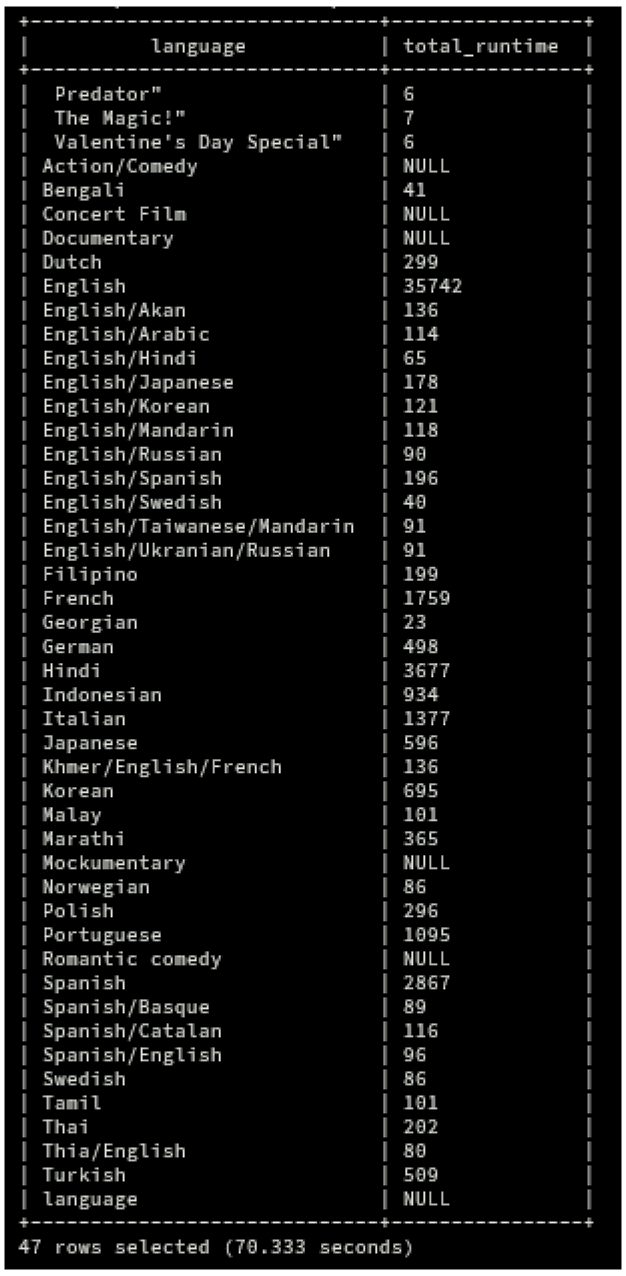
Figure 29. output average imdb score for genre

Calculate the total runtime for each language:



```
0: jdbc:hive2://> SELECT language, SUM(runtime) AS total_runtime
. . . . . . . . > FROM netflix_data
. . . . . . . . > GROUP BY language;
```

Figure 30. total runtime for each langauge

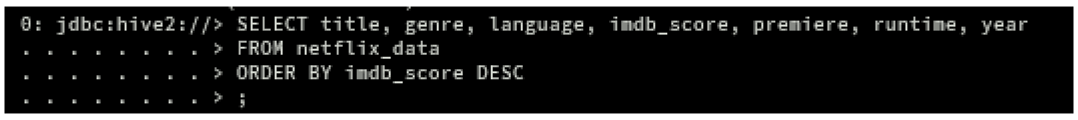


```
+-----------------------------+----------------+
|          language           | total_runtime  |
+-----------------------------+----------------+
|  Predator"                  | 6              |
|  The Magic!"                | 7              |
|  Valentine's Day Special"   | 6              |
| Action/Comedy               | NULL           |
| Bengali                     | 41             |
| Concert Film                | NULL           |
| Documentary                 | NULL           |
| Dutch                       | 299            |
| English                     | 35742          |
| English/Akan                | 136            |
| English/Arabic              | 114            |
| English/Hindi               | 65             |
| English/Japanese            | 178            |
| English/Korean              | 121            |
| English/Mandarin            | 118            |
| English/Russian             | 90             |
| English/Spanish             | 196            |
| English/Swedish             | 40             |
| English/Taiwanese/Mandarin  | 91             |
| English/Ukranian/Russian    | 91             |
| Filipino                    | 199            |
| French                      | 1759           |
| Georgian                    | 23             |
| German                      | 498            |
| Hindi                       | 3677           |
| Indonesian                  | 934            |
| Italian                     | 1377           |
| Japanese                    | 596            |
| Khmer/English/French        | 136            |
| Korean                      | 695            |
| Malay                       | 101            |
| Marathi                     | 365            |
| Mockumentary                | NULL           |
| Norwegian                   | 86             |
| Polish                      | 296            |
| Portuguese                  | 1095           |
| Romantic comedy             | NULL           |
| Spanish                     | 2867           |
| Spanish/Basque              | 89             |
| Spanish/Catalan             | 116            |
| Spanish/English             | 96             |
| Swedish                     | 86             |
| Tamil                       | 101            |
| Thai                        | 202            |
| Thia/English                | 80             |
| Turkish                     | 509            |
| language                    | NULL           |
+-----------------------------+----------------+
47 rows selected (70.333 seconds)
```

Figure 31. output runtime for each language

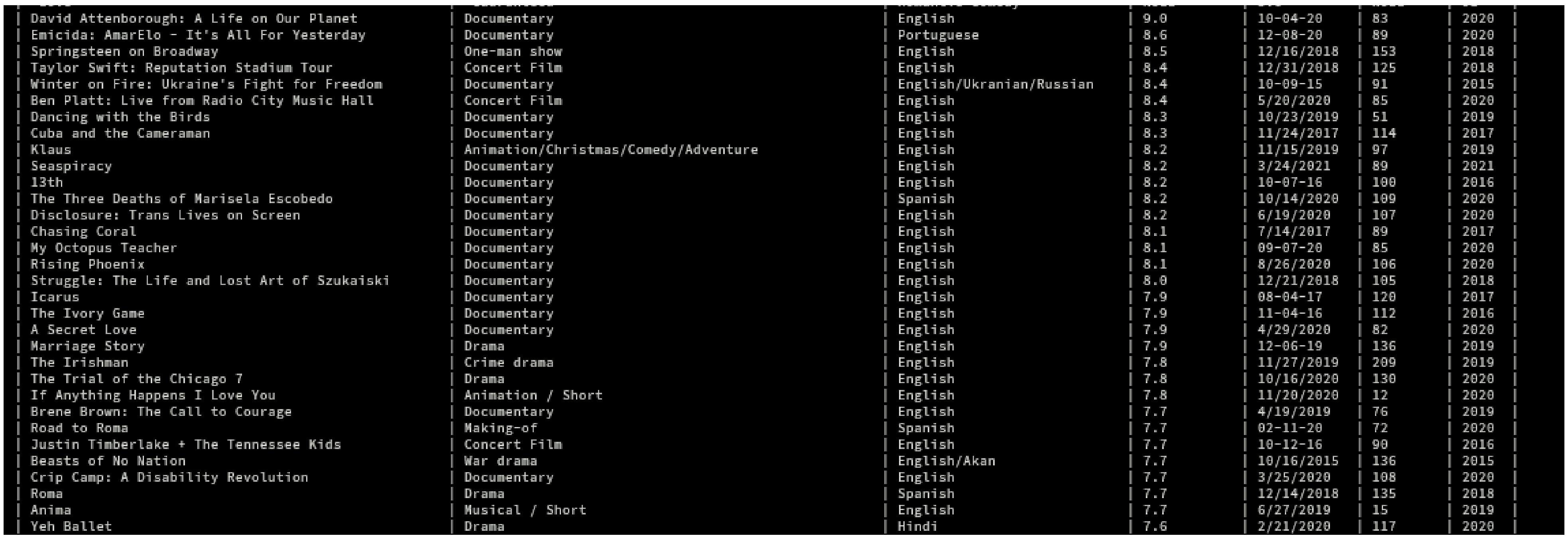
Sorting Data

Retrieve movies sorted by IMDB score in descending order:



```
0: jdbc:hive2://> SELECT title, genre, language, imdb_score, premiere, runtime, year
. . . . . . . . > FROM netflix_data
. . . . . . . . > ORDER BY imdb_score DESC
. . . . . . . . > ;
```

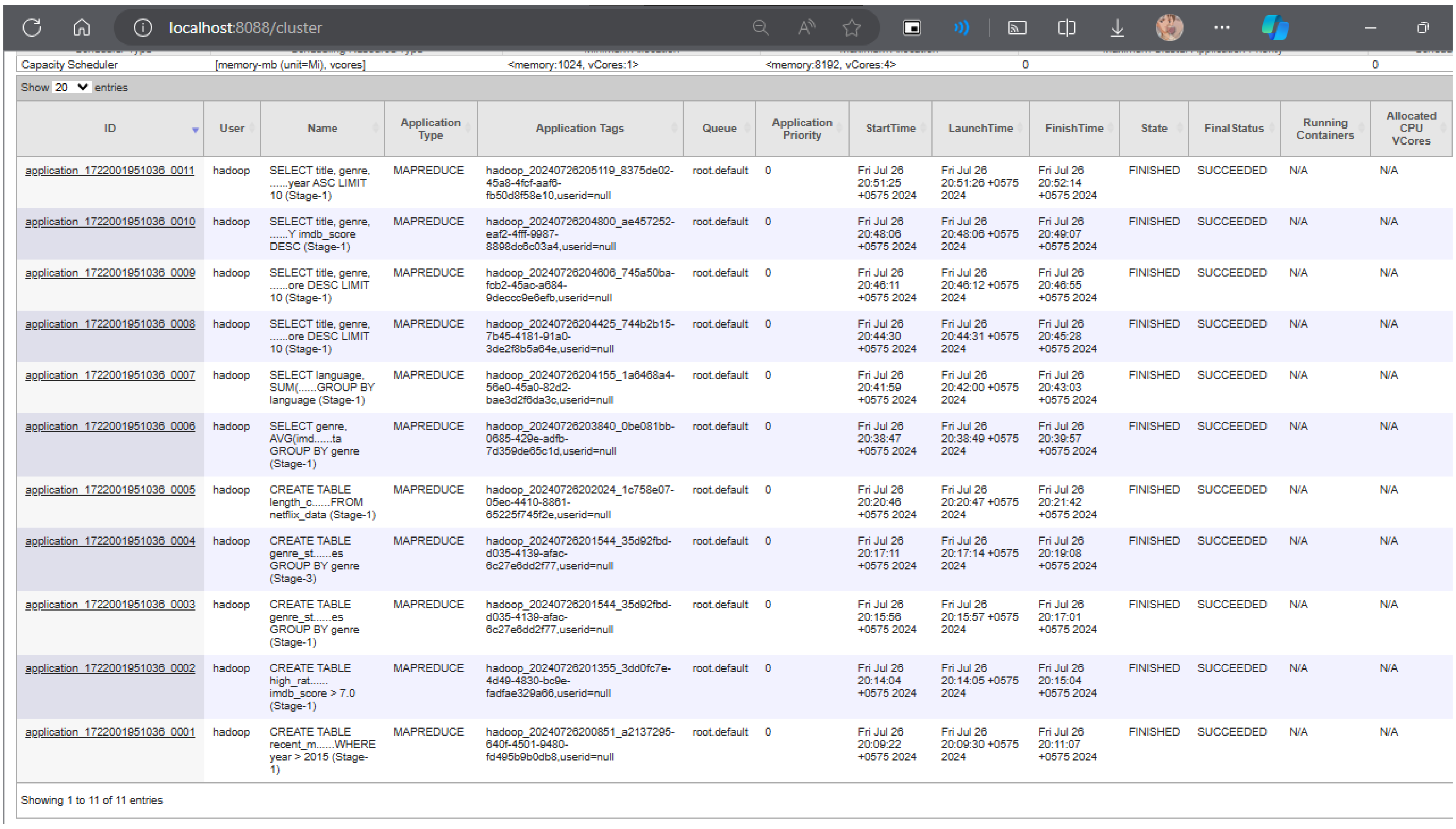
Figure 32. filter movies sorted by imdb score in decending



```
| David Attenborough: A Life on Our Planet          | Documentary                                       | English                     | 9.0      | 10-04-20    | 83        | 2020  |
| Emicida: AmarElo - It's All For Yesterday         | Documentary                                       | Portuguese                  | 8.6      | 12-08-20    | 89        | 2020  |
| Springsteen on Broadway                           | One-man show                                      | English                     | 8.5      | 12/16/2018  | 153       | 2018  |
| Taylor Swift: Reputation Stadium Tour             | Concert Film                                      | English                     | 8.4      | 12/31/2018  | 125       | 2018  |
| Winter on Fire: Ukraine's Fight for Freedom       | Documentary                                       | English/Ukranian/Russian    | 8.4      | 10-09-15    | 91        | 2015  |
| Ben Platt: Live from Radio City Music Hall        | Concert Film                                      | English                     | 8.4      | 5/20/2020   | 85        | 2020  |
| Dancing with the Birds                            | Documentary                                       | English                     | 8.3      | 10/23/2019  | 51        | 2019  |
| Cuba and the Cameraman                            | Documentary                                       | English                     | 8.3      | 11/24/2017  | 114       | 2017  |
| Klaus                                             | Animation/Christmas/Comedy/Adventure              | English                     | 8.2      | 11/15/2019  | 97        | 2019  |
| Seaspiracy                                        | Documentary                                       | English                     | 8.2      | 3/24/2021   | 89        | 2021  |
| 13th                                              | Documentary                                       | English                     | 8.2      | 10-07-16    | 100       | 2016  |
| The Three Deaths of Marisela Escobedo             | Documentary                                       | Spanish                     | 8.2      | 10/14/2020  | 109       | 2020  |
| Disclosure: Trans Lives on Screen                 | Documentary                                       | English                     | 8.2      | 6/19/2020   | 107       | 2020  |
| Chasing Coral                                     | Documentary                                       | English                     | 8.1      | 7/14/2017   | 89        | 2017  |
| My Octopus Teacher                                | Documentary                                       | English                     | 8.1      | 09-07-20    | 85        | 2020  |
| Rising Phoenix                                    | Documentary                                       | English                     | 8.1      | 8/26/2020   | 106       | 2020  |
| Struggle: The Life and Lost Art of Szukaiski      | Documentary                                       | English                     | 8.0      | 12/21/2018  | 105       | 2018  |
| Icarus                                            | Documentary                                       | English                     | 7.9      | 08-04-17    | 120       | 2017  |
| The Ivory Game                                    | Documentary                                       | English                     | 7.9      | 11-04-16    | 112       | 2016  |
| A Secret Love                                     | Documentary                                       | English                     | 7.9      | 4/29/2020   | 82        | 2020  |
| Marriage Story                                    | Drama                                             | English                     | 7.9      | 12-06-19    | 136       | 2019  |
| The Irishman                                      | Crime drama                                       | English                     | 7.8      | 11/27/2019  | 209       | 2019  |
| The Trial of the Chicago 7                        | Drama                                             | English                     | 7.8      | 10/16/2020  | 130       | 2020  |
| If Anything Happens I Love You                    | Animation / Short                                 | English                     | 7.8      | 11/20/2020  | 12        | 2020  |
| Brene Brown: The Call to Courage                  | Documentary                                       | English                     | 7.7      | 4/19/2019   | 76        | 2019  |
| Road to Roma                                      | Making-of                                         | Spanish                     | 7.7      | 02-11-20    | 72        | 2020  |
| Justin Timberlake + The Tennessee Kids            | Concert Film                                      | English                     | 7.7      | 10-12-16    | 90        | 2016  |
| Beasts of No Nation                               | War drama                                         | English/Akan                | 7.7      | 10/16/2015  | 136       | 2015  |
| Crip Camp: A Disability Revolution                | Documentary                                       | English                     | 7.7      | 3/25/2020   | 108       | 2020  |
| Roma                                              | Drama                                             | Spanish                     | 7.7      | 12/14/2018  | 135       | 2018  |
| Anima                                             | Musical / Short                                   | English                     | 7.7      | 6/27/2019   | 15        | 2019  |
| Yeh Ballet                                        | Drama                                             | Hindi                       | 7.6      | 2/21/2020   | 117       | 2020  |
```

Figure 33. output movies sorted by imdb score in decending

map reduce and other all application on hadoop cluster in localhost 8088



Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0	0

Show 20 entries

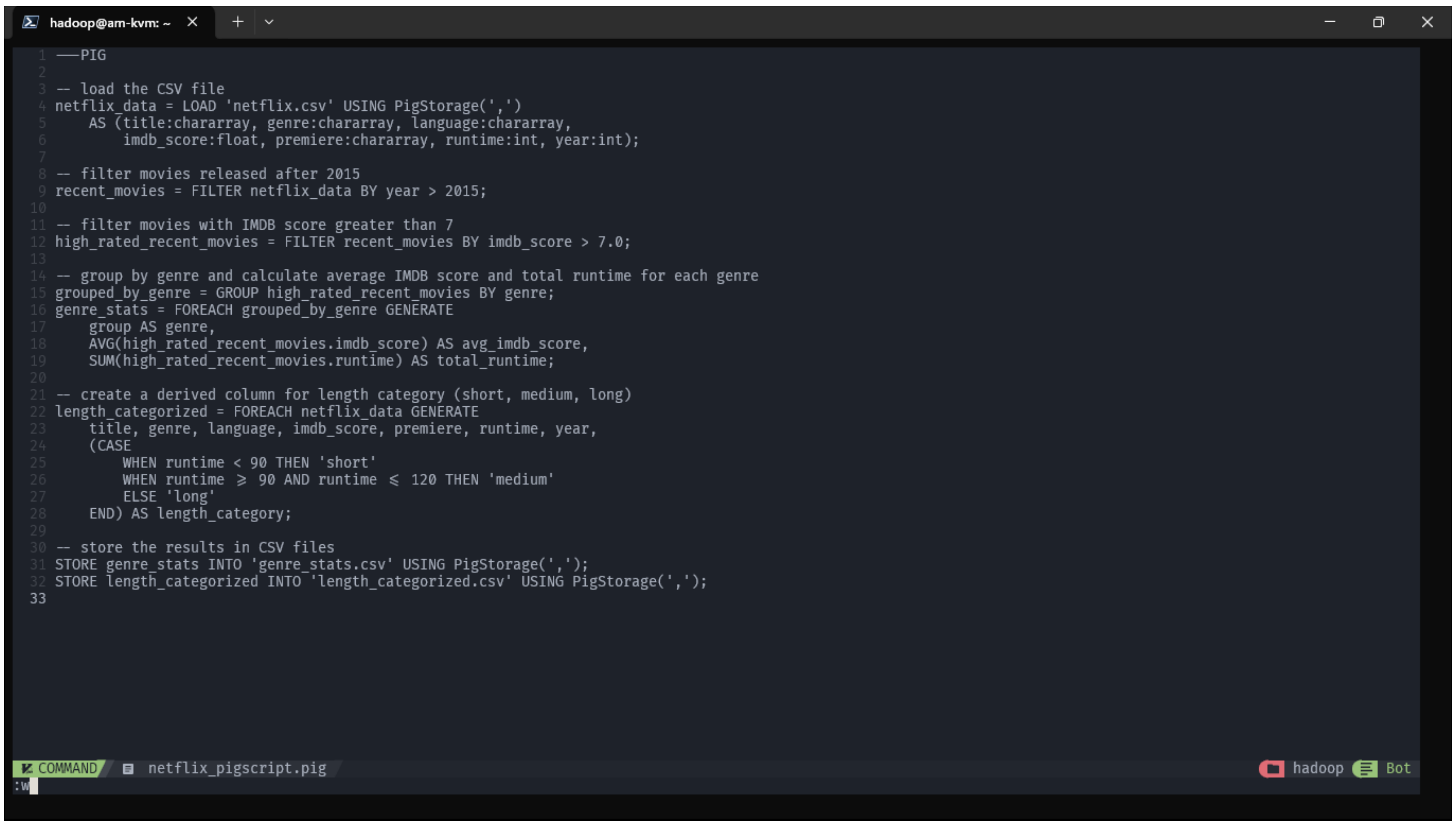
ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores
application_1722001951036_0011	hadoop	SELECT title, genre, ...year ASC LIMIT 10 (Stage-1)	MAPREDUCE	hadoop_20240726205119_8375de02-45a8-4fcf-aaf6-fb50d8f58e10,userid=null	root.default	0	Fri Jul 26 20:51:25 +0575 2024	Fri Jul 26 20:51:26 +0575 2024	Fri Jul 26 20:52:14 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0010	hadoop	SELECT title, genre, ...Y imdb_score DESC (Stage-1)	MAPREDUCE	hadoop_20240726204800_ae457252-eaf2-4fff-9987-8898dc6c03a4,userid=null	root.default	0	Fri Jul 26 20:48:06 +0575 2024	Fri Jul 26 20:48:06 +0575 2024	Fri Jul 26 20:49:07 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0009	hadoop	SELECT title, genre, ...ore DESC LIMIT 10 (Stage-1)	MAPREDUCE	hadoop_20240726204606_745a50ba-fcb2-45ac-a684-9deccc9e6efb,userid=null	root.default	0	Fri Jul 26 20:46:11 +0575 2024	Fri Jul 26 20:46:12 +0575 2024	Fri Jul 26 20:46:55 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0008	hadoop	SELECT title, genre, ...ore DESC LIMIT 10 (Stage-1)	MAPREDUCE	hadoop_20240726204425_744b2b15-7b45-4181-91a0-3de2f8b5a64e,userid=null	root.default	0	Fri Jul 26 20:44:30 +0575 2024	Fri Jul 26 20:44:31 +0575 2024	Fri Jul 26 20:45:28 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0007	hadoop	SELECT language, SUM(......GROUP BY language (Stage-1)	MAPREDUCE	hadoop_20240726204155_1a6468a4-56e0-45a0-82d2-bae3d2f6da3c,userid=null	root.default	0	Fri Jul 26 20:41:59 +0575 2024	Fri Jul 26 20:42:00 +0575 2024	Fri Jul 26 20:43:03 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0006	hadoop	SELECT genre, AVG(imd......ta GROUP BY genre (Stage-1)	MAPREDUCE	hadoop_20240726203840_0be081bb-0685-429e-adfb-7d359de65c1d,userid=null	root.default	0	Fri Jul 26 20:38:47 +0575 2024	Fri Jul 26 20:38:49 +0575 2024	Fri Jul 26 20:39:57 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0005	hadoop	CREATE TABLE length_c......FROM netflix_data (Stage-1)	MAPREDUCE	hadoop_20240726202024_1c758e07-05ec-4410-8861-65225f745f2e,userid=null	root.default	0	Fri Jul 26 20:20:46 +0575 2024	Fri Jul 26 20:20:47 +0575 2024	Fri Jul 26 20:21:42 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0004	hadoop	CREATE TABLE genre_st......es GROUP BY genre (Stage-3)	MAPREDUCE	hadoop_20240726201544_35d92fbd-d035-4139-afac-6c27e6dd2f77,userid=null	root.default	0	Fri Jul 26 20:17:11 +0575 2024	Fri Jul 26 20:17:14 +0575 2024	Fri Jul 26 20:19:08 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0003	hadoop	CREATE TABLE genre_st......es GROUP BY genre (Stage-1)	MAPREDUCE	hadoop_20240726201544_35d92fbd-d035-4139-afac-6c27e6dd2f77,userid=null	root.default	0	Fri Jul 26 20:15:56 +0575 2024	Fri Jul 26 20:15:57 +0575 2024	Fri Jul 26 20:17:01 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0002	hadoop	CREATE TABLE high_rat...... imdb_score > 7.0 (Stage-1)	MAPREDUCE	hadoop_20240726201355_3dd0fc7e-4d49-4830-bc9e-fadfae329a66,userid=null	root.default	0	Fri Jul 26 20:14:04 +0575 2024	Fri Jul 26 20:14:05 +0575 2024	Fri Jul 26 20:15:04 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A
application_1722001951036_0001	hadoop	CREATE TABLE recent_m......WHERE year > 2015 (Stage-1)	MAPREDUCE	hadoop_20240726200851_a2137295-640f-4501-9480-fd495b9b0db8,userid=null	root.default	0	Fri Jul 26 20:09:22 +0575 2024	Fri Jul 26 20:09:30 +0575 2024	Fri Jul 26 20:11:07 +0575 2024	FINISHED	SUCCEEDED	N/A	N/A

Showing 1 to 11 of 11 entries

Figure 34. hadoop localhost

7.3. Pig

The Entire Script

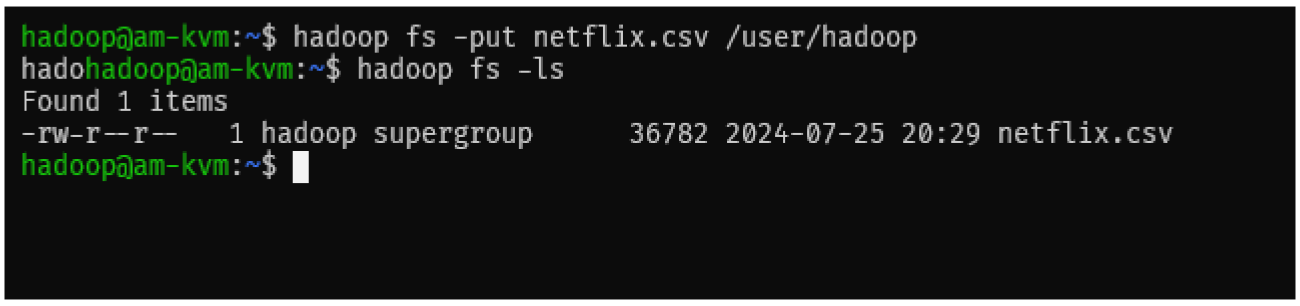


```
——PIG

-- load the CSV file
netflix_data = LOAD 'netflix.csv' USING PigStorage(',')
    AS (title:chararray, genre:chararray, language:chararray,
        imdb_score:float, premiere:chararray, runtime:int, year:int);

-- filter movies released after 2015
recent_movies = FILTER netflix_data BY year > 2015;

-- filter movies with IMDB score greater than 7
high_rated_recent_movies = FILTER recent_movies BY imdb_score > 7.0;

-- group by genre and calculate average IMDB score and total runtime for each genre
grouped_by_genre = GROUP high_rated_recent_movies BY genre;
genre_stats = FOREACH grouped_by_genre GENERATE
    group AS genre,
    AVG(high_rated_recent_movies.imdb_score) AS avg_imdb_score,
    SUM(high_rated_recent_movies.runtime) AS total_runtime;

-- create a derived column for length category (short, medium, long)
length_categorized = FOREACH netflix_data GENERATE
    title, genre, language, imdb_score, premiere, runtime, year,
    (CASE
        WHEN runtime < 90 THEN 'short'
        WHEN runtime ⩾ 90 AND runtime ⩽ 120 THEN 'medium'
        ELSE 'long'
    END) AS length_category;

-- store the results in CSV files
STORE genre_stats INTO 'genre_stats.csv' USING PigStorage(',');
STORE length_categorized INTO 'length_categorized.csv' USING PigStorage(',');

:w
```

Figure 35. pig script

Move the dataset from host to hdfs

In order to perform queries on the dataset, we need to move the dataset from the host machine to the hadoop file system (hdfs). We use the hadoop fs -put command to copy the file into the hadoop file system.

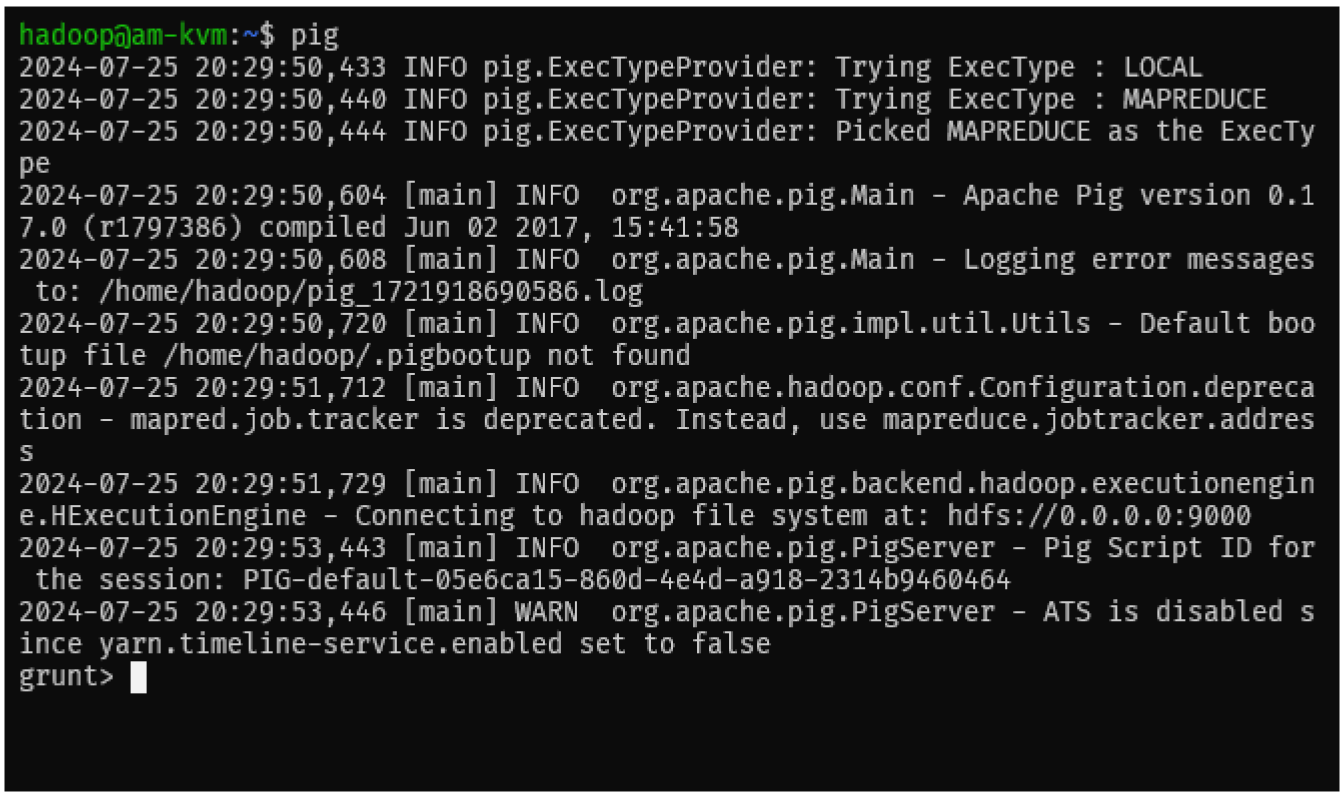


```
hadoop@am-kvm:~$ hadoop fs -put netflix.csv /user/hadoop
hadohadoop@am-kvm:~$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 hadoop supergroup      36782 2024-07-25 20:29 netflix.csv
hadoop@am-kvm:~$
```

Figure 36. put Netflix dataset in hadoop dir

Pig Interface

Now, we can go into the pig interface to query the dataset. We can run the pig command for this.

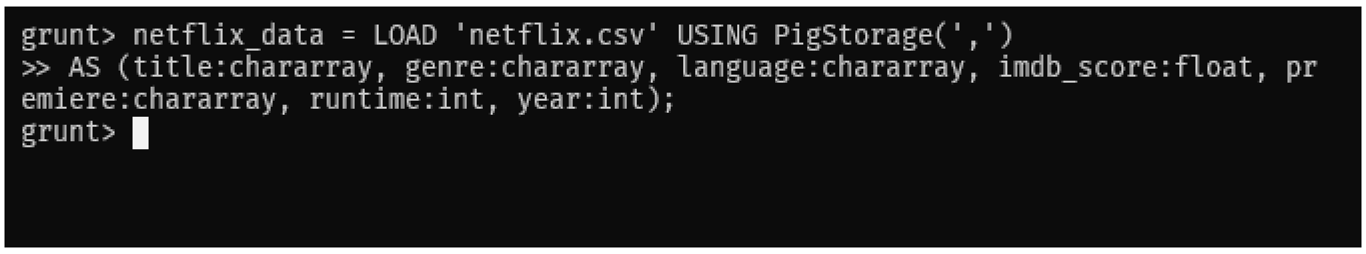


```
hadoop@am-kvm:~$ pig
2024-07-25 20:29:50,433 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-07-25 20:29:50,440 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-07-25 20:29:50,444 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecTy
pe
2024-07-25 20:29:50,604 [main] INFO  org.apache.pig.Main - Apache Pig version 0.1
7.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-07-25 20:29:50,608 [main] INFO  org.apache.pig.Main - Logging error messages
 to: /home/hadoop/pig_1721918690586.log
2024-07-25 20:29:50,720 [main] INFO  org.apache.pig.impl.util.Utils - Default boo
tup file /home/hadoop/.pigbootup not found
2024-07-25 20:29:51,712 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addres
s
2024-07-25 20:29:51,729 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.HExecutionEngine - Connecting to hadoop file system at: hdfs://0.0.0.0:9000
2024-07-25 20:29:53,443 [main] INFO  org.apache.pig.PigServer - Pig Script ID for
 the session: PIG-default-05e6ca15-860d-4e4d-a918-2314b9460464
2024-07-25 20:29:53,446 [main] WARN  org.apache.pig.PigServer - ATS is disabled s
ince yarn.timeline-service.enabled set to false
grunt>
```

Figure 37. pig shell

Load the dataset

The following query allows us to load the netflix dataset, specifically, a CSV file named ‘netflix.csv’. The “USING PigStorage(‘,’)” function specifies that the data is comma-separated. The “AS” clause in this query help us define the schema of the dataset, specifying what columns the dataset includes and what dataset those columns use.



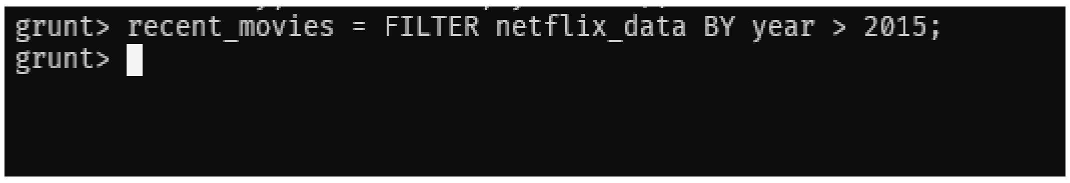
```
grunt> netflix_data = LOAD 'netflix.csv' USING PigStorage(',')
>> AS (title:chararray, genre:chararray, language:chararray, imdb_score:float, pr
emiere:chararray, runtime:int, year:int);
grunt> 
```

Figure 38. load csv to pig

Filtering

Filtering movies released after 2015

This query filters the “netflix_data” to include only movies released after 2015 and stores it in “recent_movies”.

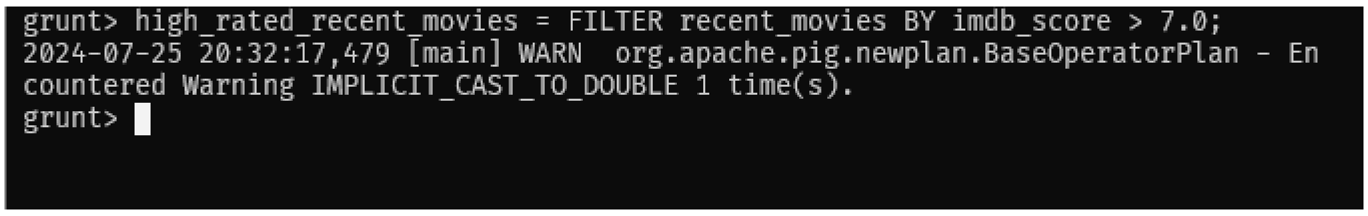


```
grunt> recent_movies = FILTER netflix_data BY year > 2015;
grunt> 
```

Figure 39. filter movies after 2025

Filtering movies with IMDb rating more than 7

This query filters “recent_movies” to include movies that have a IMDb score greater than 7.

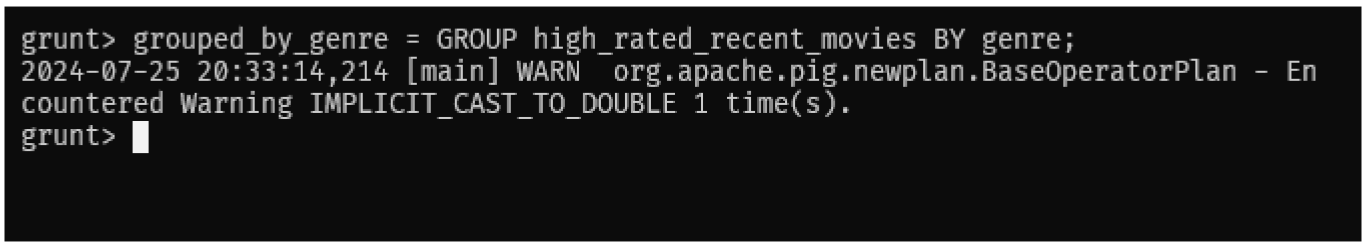


```
grunt> high_rated_recent_movies = FILTER recent_movies BY imdb_score > 7.0;
2024-07-25 20:32:17,479 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> 
```

Figure 40. filter movie with more than 7 score

Group by genre

This line groups the ‘high_rated_recent_movies’ dataset by the genre field.

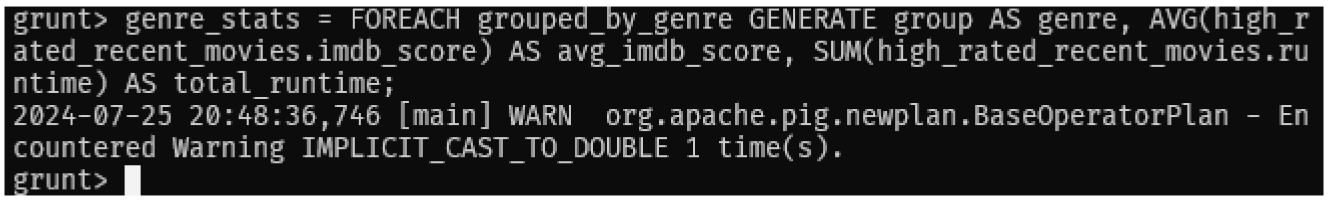


```
grunt> grouped_by_genre = GROUP high_rated_recent_movies BY genre;
2024-07-25 20:33:14,214 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> 
```

Figure 41. group movie by genre

Get genre stats

This query iterates over each group created by the GROUP statement in the previous query. The group field is renamed as genre, the second field is the average IMDb score of each genre, and the third field calculates the total runtime of all movies within each genre.

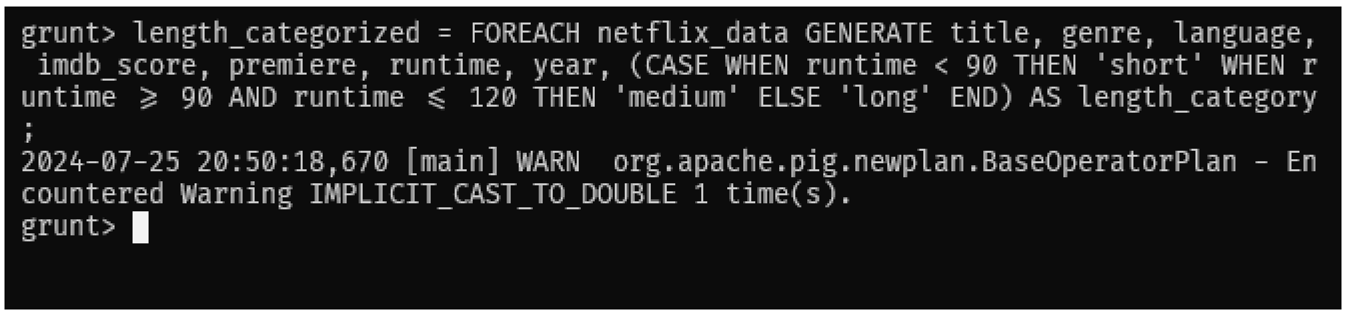


```
grunt> genre_stats = FOREACH grouped_by_genre GENERATE group AS genre, AVG(high_r
ated_recent_movies.imdb_score) AS avg_imdb_score, SUM(high_rated_recent_movies.ru
ntime) AS total_runtime;
2024-07-25 20:48:36,746 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> 
```

Figure 42. group genre stats

Create a derived column for length category

This query iterates over each record in the “netflix_data” dataset, and creates a new derived column called “length_category”. The “length_category” column assigns enumerated values of “short” if the runtime of the show is less than 90 minutes, “medium” if runtime is less than or equal to 120 minutes, and “long” if otherwise

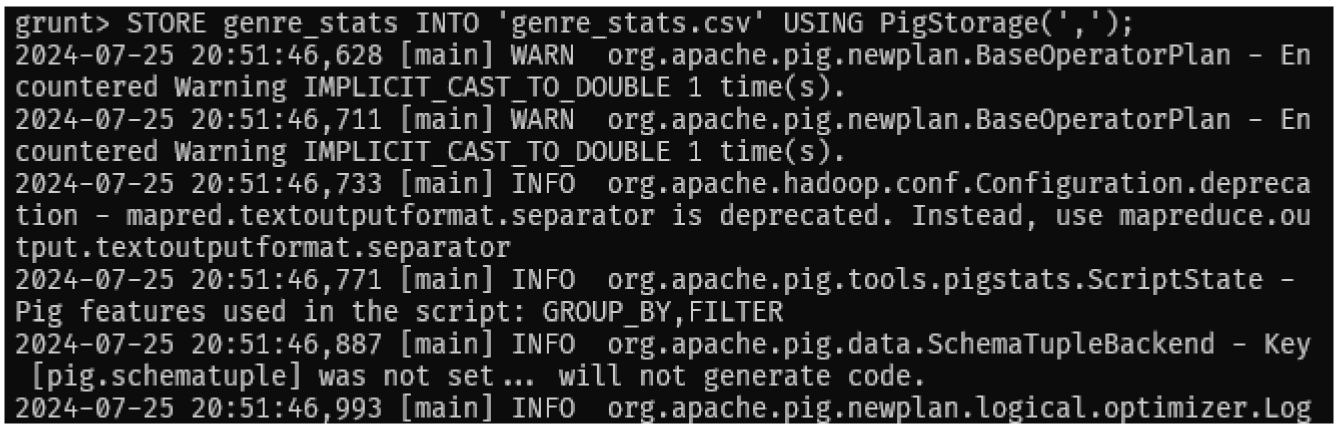


```
grunt> length_categorized = FOREACH netflix_data GENERATE title, genre, language,
 imdb_score, premiere, runtime, year, (CASE WHEN runtime < 90 THEN 'short' WHEN r
untime ≥ 90 AND runtime ≤ 120 THEN 'medium' ELSE 'long' END) AS length_category
;
2024-07-25 20:50:18,670 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> █
```

Figure 43. derive column for length category

For readability: Store genre stats

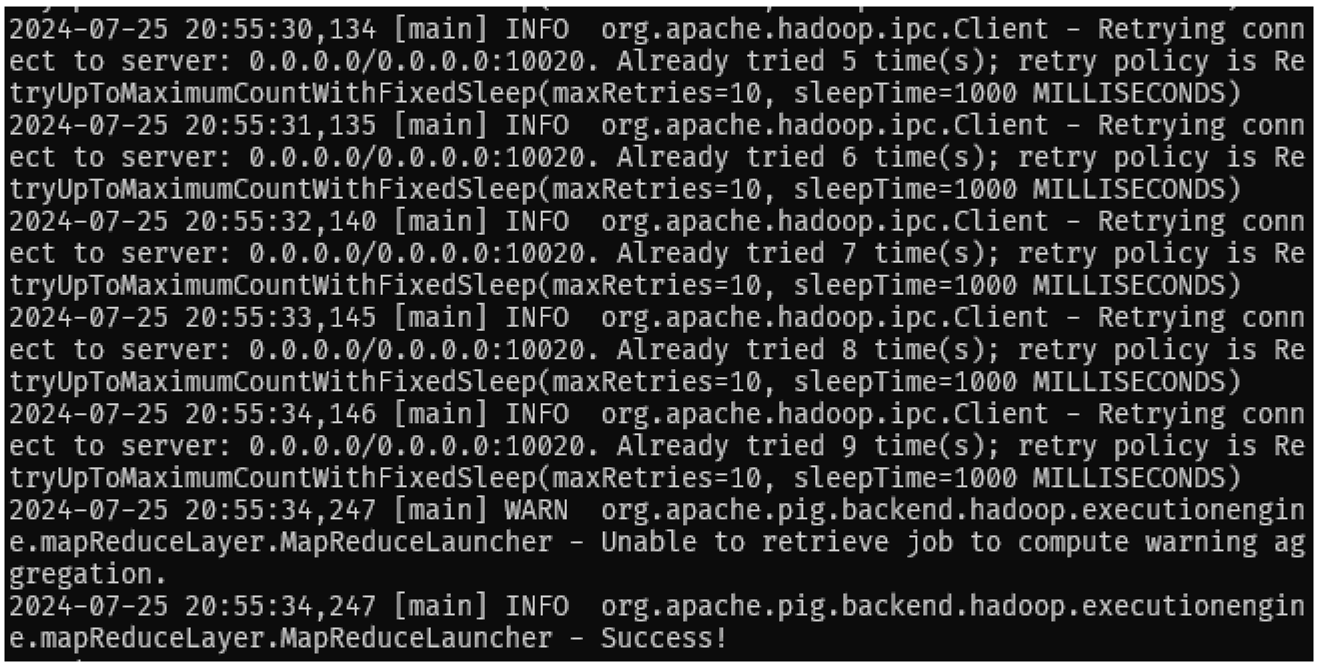
We can store the output of the queries performed above into a file. By using the STORE command, we can write the dataset to a file. Here, this query saves the ‘genre_stats’ dataset into a CSV file named ‘genre_stats.csv’ using commas (‘,’) as a delimiter.



```
grunt> STORE genre_stats INTO 'genre_stats.csv' USING PigStorage(',');
2024-07-25 20:51:46,628 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2024-07-25 20:51:46,711 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - En
countered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2024-07-25 20:51:46,733 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.ou
tput.textoutputformat.separator
2024-07-25 20:51:46,771 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: GROUP_BY,FILTER
2024-07-25 20:51:46,887 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key
 [pig.schematuple] was not set... will not generate code.
2024-07-25 20:51:46,993 [main] INFO  org.apache.pig.newplan.logical.optimizer.Log
```

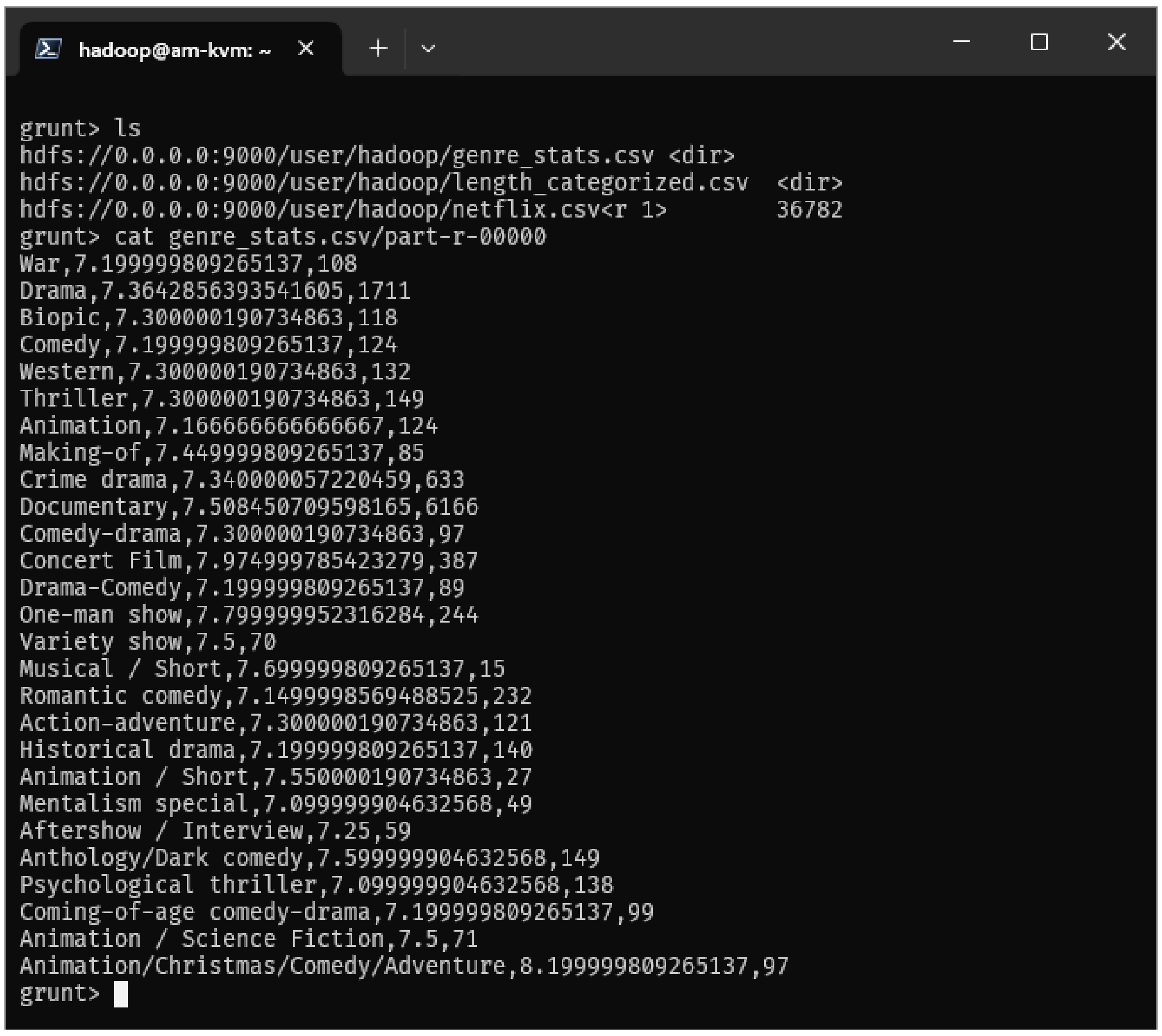
Figure 44. for readability and genre stats

Output:



```
2024-07-25 20:55:30,134 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 20:55:31,135 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 20:55:32,140 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 20:55:33,145 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 20:55:34,146 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 20:55:34,247 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning ag
gregation.
2024-07-25 20:55:34,247 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
```

Figure 45. MR process of genre stats

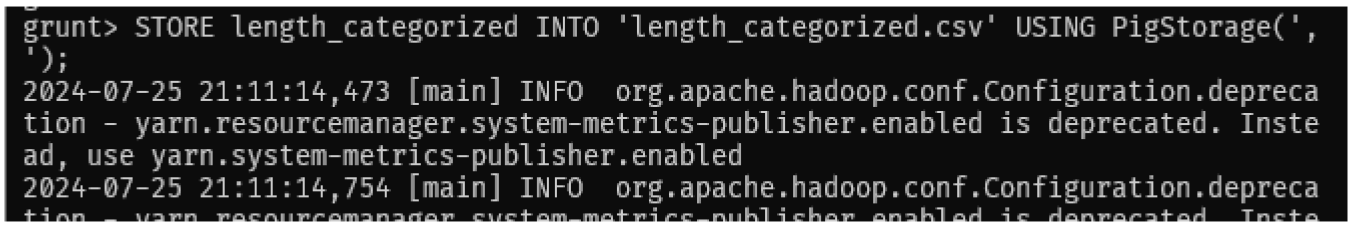


```
grunt> ls
hdfs://0.0.0.0:9000/user/hadoop/genre_stats.csv <dir>
hdfs://0.0.0.0:9000/user/hadoop/length_categorized.csv  <dir>
hdfs://0.0.0.0:9000/user/hadoop/netflix.csv<r 1>        36782
grunt> cat genre_stats.csv/part-r-00000
War,7.199999809265137,108
Drama,7.3642856393541605,1711
Biopic,7.300000190734863,118
Comedy,7.199999809265137,124
Western,7.300000190734863,132
Thriller,7.300000190734863,149
Animation,7.166666666666667,124
Making-of,7.449999809265137,85
Crime drama,7.340000057220459,633
Documentary,7.508450709598165,6166
Comedy-drama,7.300000190734863,97
Concert Film,7.974999785423279,387
Drama-Comedy,7.199999809265137,89
One-man show,7.799999952316284,244
Variety show,7.5,70
Musical / Short,7.699999809265137,15
Romantic comedy,7.1499998569488525,232
Action-adventure,7.300000190734863,121
Historical drama,7.199999809265137,140
Animation / Short,7.550000190734863,27
Mentalism special,7.099999904632568,49
Aftershow / Interview,7.25,59
Anthology/Dark comedy,7.599999904632568,149
Psychological thriller,7.099999904632568,138
Coming-of-age comedy-drama,7.199999809265137,99
Animation / Science Fiction,7.5,71
Animation/Christmas/Comedy/Adventure,8.199999809265137,97
grunt> 
```

Figure 46. output of genre stats

Store length categories

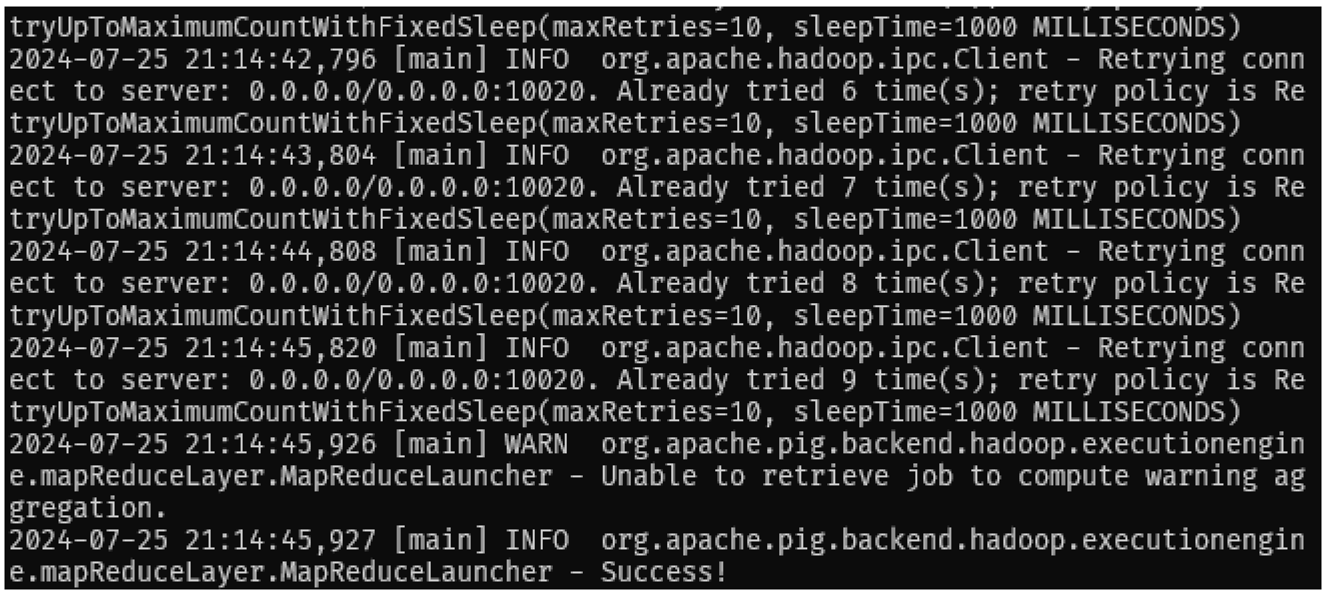
Similar to the previous query, this query also write the output of a dataset into a file. This query saves the ‘length_categorized’ dataset into a CSV file ‘length_categorized.csv’, using commas (‘,’) as delimiters.



```
grunt> STORE length_categorized INTO 'length_categorized.csv' USING PigStorage(',
');
2024-07-25 21:11:14,473 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inste
ad, use yarn.system-metrics-publisher.enabled
2024-07-25 21:11:14,754 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inste
```

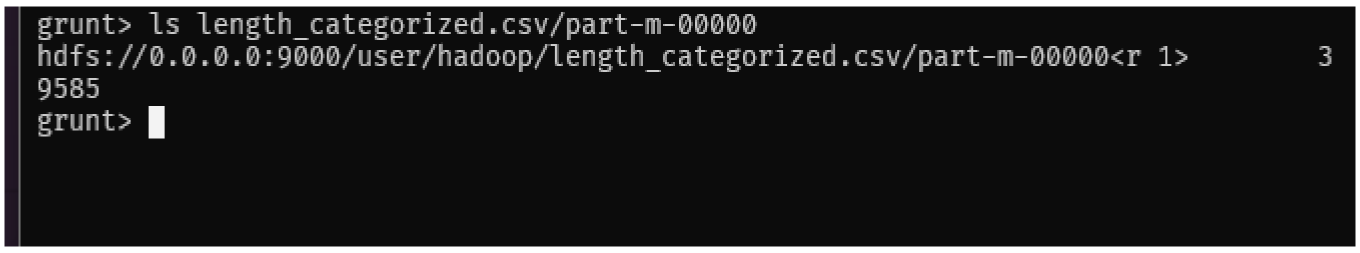
Figure 47. store length of categories

Output:



```
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 21:14:42,796 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 21:14:43,804 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 21:14:44,808 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 21:14:45,820 [main] INFO  org.apache.hadoop.ipc.Client - Retrying conn
ect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is Re
tryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-07-25 21:14:45,926 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning ag
gregation.
2024-07-25 21:14:45,927 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
```

Figure 48. map reduce process of length of categories

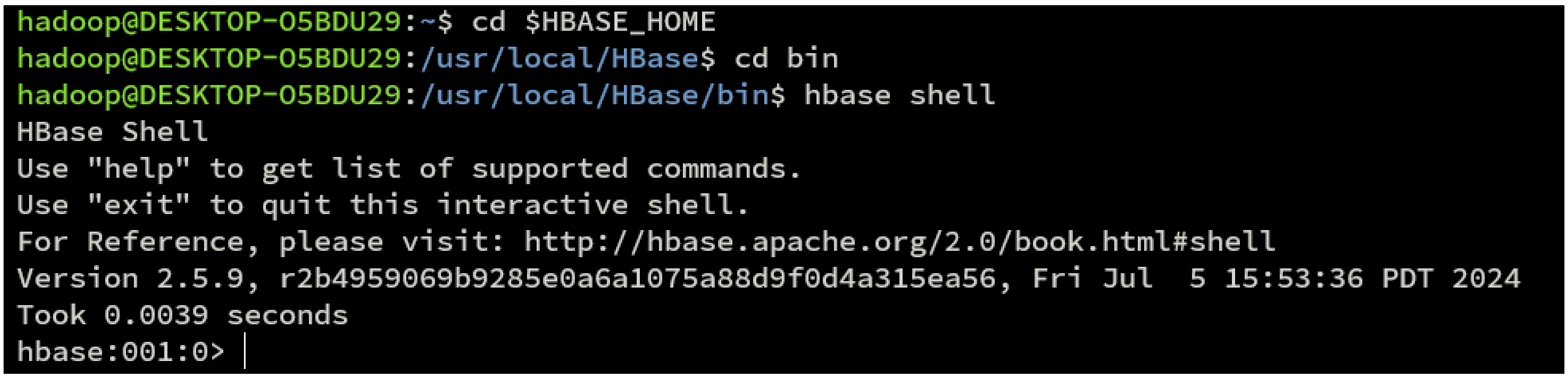


```
grunt> ls length_categorized.csv/part-m-00000
hdfs://0.0.0.0:9000/user/hadoop/length_categorized.csv/part-m-00000<r 1>        3
9585
grunt> 
```

Figure 49. output of length of categories

7.4. HBase

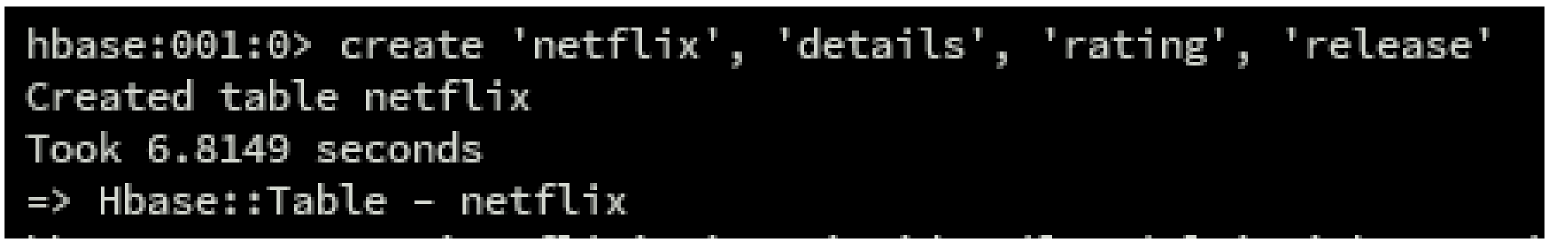
open hbase shell



```
hadoop@DESKTOP-05BDU29:~$ cd $HBASE_HOME
hadoop@DESKTOP-05BDU29:/usr/local/HBase$ cd bin
hadoop@DESKTOP-05BDU29:/usr/local/HBase/bin$ hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.5.9, r2b4959069b9285e0a6a1075a88d9f0d4a315ea56, Fri Jul  5 15:53:36 PDT 2024
Took 0.0039 seconds
hbase:001:0> |
```

Figure 50. hbase shell

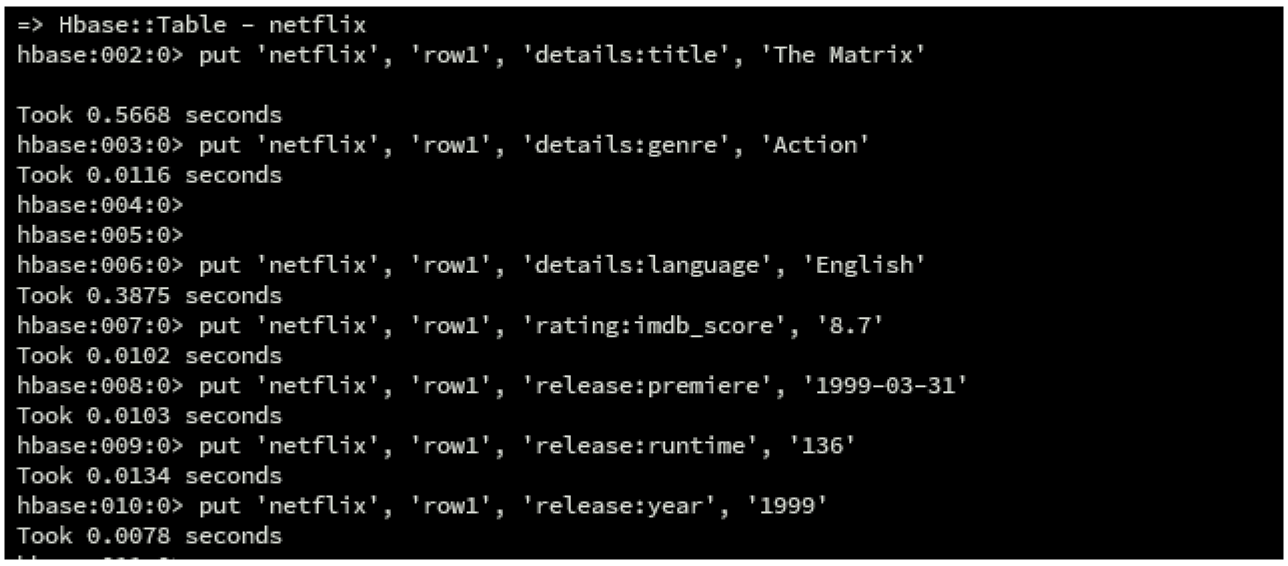
create table



```
hbase:001:0> create 'netflix', 'details', 'rating', 'release'
Created table netflix
Took 6.8149 seconds
=> Hbase::Table - netflix
```

Figure 51. create table

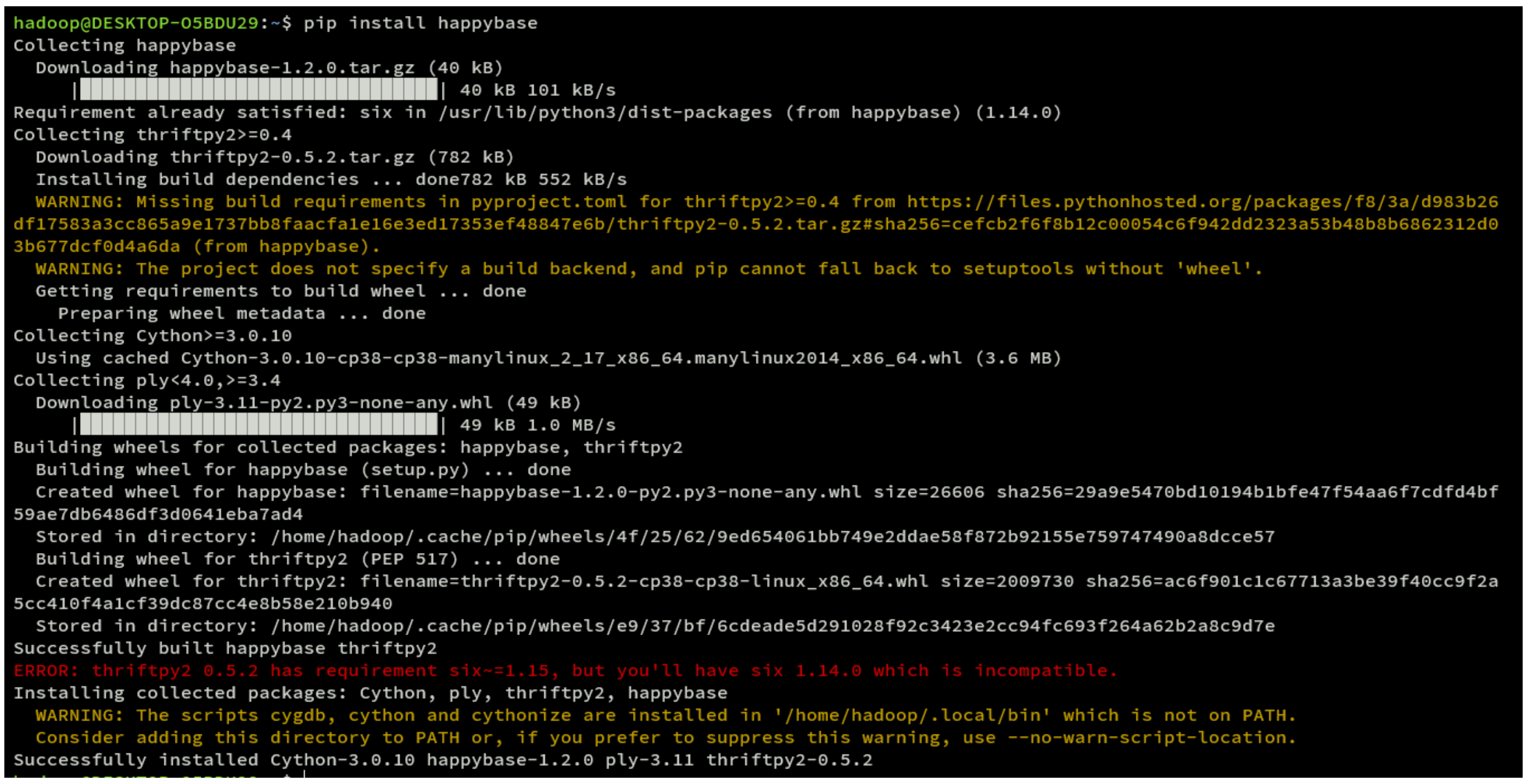
insert data row by row



```
=> Hbase::Table - netflix
hbase:002:0> put 'netflix', 'row1', 'details:title', 'The Matrix'

Took 0.5668 seconds
hbase:003:0> put 'netflix', 'row1', 'details:genre', 'Action'
Took 0.0116 seconds
hbase:004:0>
hbase:005:0>
hbase:006:0> put 'netflix', 'row1', 'details:language', 'English'
Took 0.3875 seconds
hbase:007:0> put 'netflix', 'row1', 'rating:imdb_score', '8.7'
Took 0.0102 seconds
hbase:008:0> put 'netflix', 'row1', 'release:premiere', '1999-03-31'
Took 0.0103 seconds
hbase:009:0> put 'netflix', 'row1', 'release:runtime', '136'
Took 0.0134 seconds
hbase:010:0> put 'netflix', 'row1', 'release:year', '1999'
Took 0.0078 seconds
```

Figure 52. insert data row by row

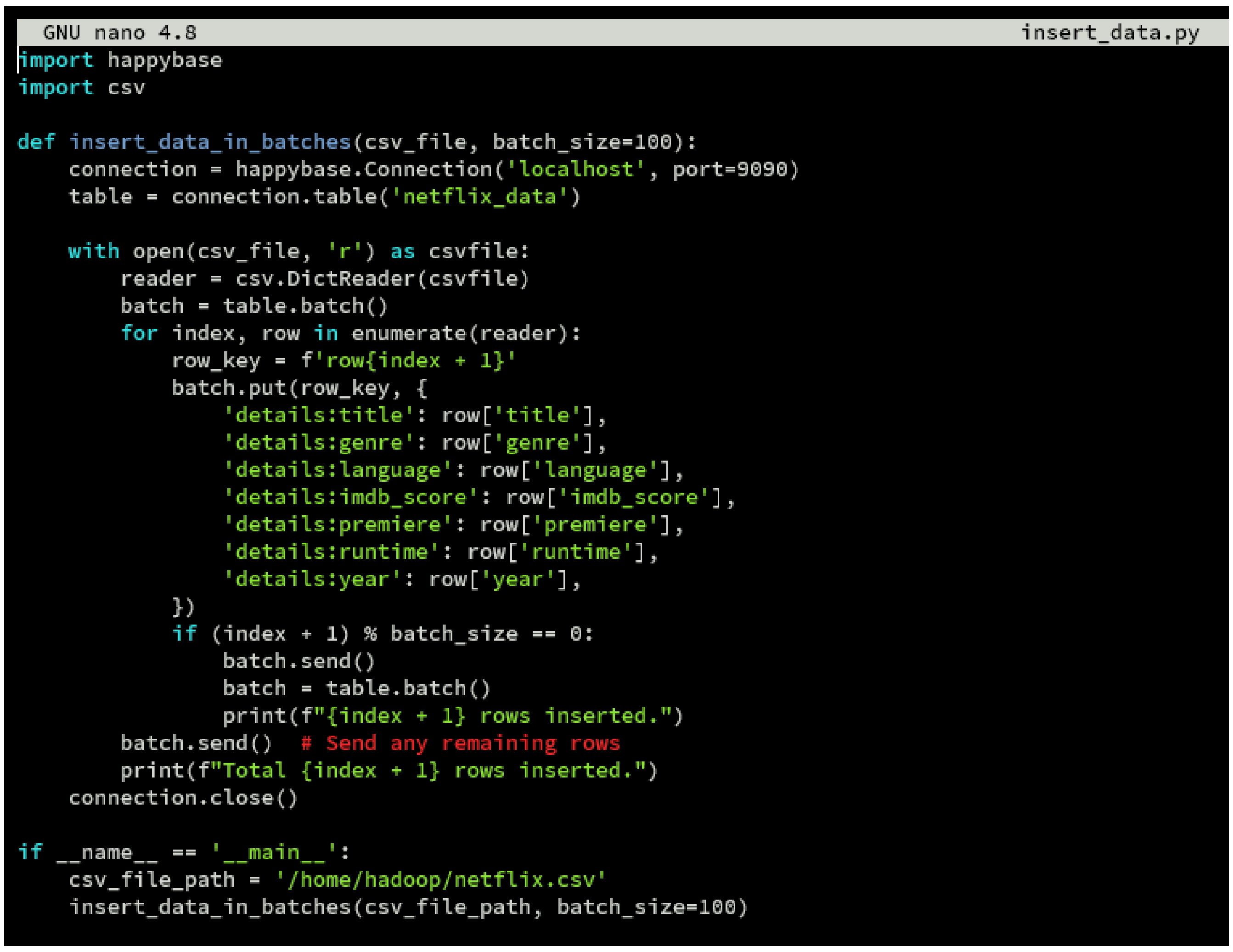
install happy base to automate insertion process in python script



```
hadoop@DESKTOP-05BDU29:~$ pip install happybase
Collecting happybase
  Downloading happybase-1.2.0.tar.gz (40 kB)
     |████████████████████████████████| 40 kB 101 kB/s
Requirement already satisfied: six in /usr/lib/python3/dist-packages (from happybase) (1.14.0)
Collecting thriftpy2>=0.4
  Downloading thriftpy2-0.5.2.tar.gz (782 kB)
  Installing build dependencies ... done782 kB 552 kB/s
  WARNING: Missing build requirements in pyproject.toml for thriftpy2>=0.4 from https://files.pythonhosted.org/packages/f8/3a/d983b26
df17583a3cc865a9e1737bb8faacfa1e16e3ed17353ef48847e6b/thriftpy2-0.5.2.tar.gz#sha256=cefcb2f6f8b12c00054c6f942dd2323a53b48b8b6862312d0
3b677dcf0d4a6da (from happybase).
  WARNING: The project does not specify a build backend, and pip cannot fall back to setuptools without 'wheel'.
  Getting requirements to build wheel ... done
    Preparing wheel metadata ... done
Collecting Cython>=3.0.10
  Using cached Cython-3.0.10-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.6 MB)
Collecting ply<4.0,>=3.4
  Downloading ply-3.11-py2.py3-none-any.whl (49 kB)
     |████████████████████████████████| 49 kB 1.0 MB/s
Building wheels for collected packages: happybase, thriftpy2
  Building wheel for happybase (setup.py) ... done
  Created wheel for happybase: filename=happybase-1.2.0-py2.py3-none-any.whl size=26606 sha256=29a9e5470bd10194b1bfe47f54aa6f7cdfd4bf
59ae7db6486df3d0641eba7ad4
  Stored in directory: /home/hadoop/.cache/pip/wheels/4f/25/62/9ed654061bb749e2ddae58f872b92155e759747490a8dcce57
  Building wheel for thriftpy2 (PEP 517) ... done
  Created wheel for thriftpy2: filename=thriftpy2-0.5.2-cp38-cp38-linux_x86_64.whl size=2009730 sha256=ac6f901c1c67713a3be39f40cc9f2a
5cc410f4a1cf39dc87cc4e8b58e210b940
  Stored in directory: /home/hadoop/.cache/pip/wheels/e9/37/bf/6cdeade5d291028f92c3423e2cc94fc693f264a62b2a8c9d7e
Successfully built happybase thriftpy2
ERROR: thriftpy2 0.5.2 has requirement six~=1.15, but you'll have six 1.14.0 which is incompatible.
Installing collected packages: Cython, ply, thriftpy2, happybase
  WARNING: The scripts cygdb, cython and cythonize are installed in '/home/hadoop/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed Cython-3.0.10 happybase-1.2.0 ply-3.11 thriftpy2-0.5.2
```

Figure 53. python happy base pip i

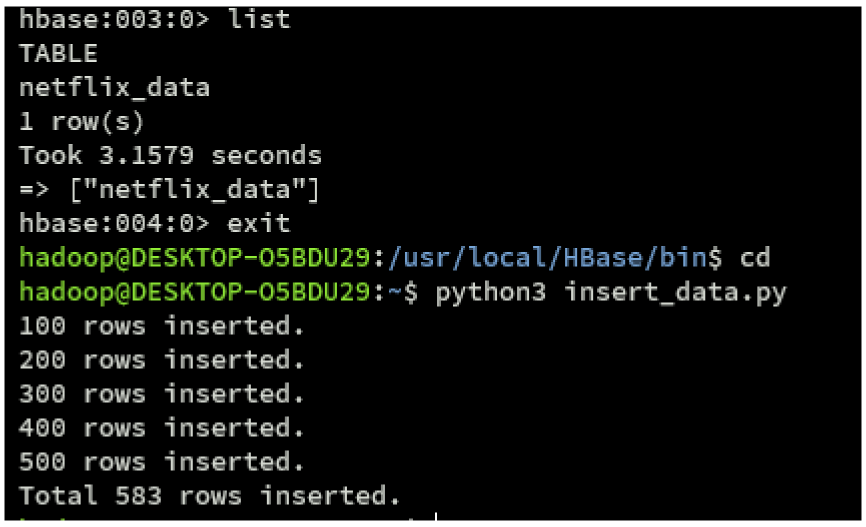
nstall



```
  GNU nano 4.8                                                        insert_data.py
import happybase
import csv

def insert_data_in_batches(csv_file, batch_size=100):
    connection = happybase.Connection('localhost', port=9090)
    table = connection.table('netflix_data')

    with open(csv_file, 'r') as csvfile:
        reader = csv.DictReader(csvfile)
        batch = table.batch()
        for index, row in enumerate(reader):
            row_key = f'row{index + 1}'
            batch.put(row_key, {
                'details:title': row['title'],
                'details:genre': row['genre'],
                'details:language': row['language'],
                'details:imdb_score': row['imdb_score'],
                'details:premiere': row['premiere'],
                'details:runtime': row['runtime'],
                'details:year': row['year'],
            })
            if (index + 1) % batch_size == 0:
                batch.send()
                batch = table.batch()
                print(f"{index + 1} rows inserted.")
        batch.send()  # Send any remaining rows
        print(f"Total {index + 1} rows inserted.")
    connection.close()

if __name__ == '__main__':
    csv_file_path = '/home/hadoop/netflix.csv'
    insert_data_in_batches(csv_file_path, batch_size=100)
```

Figure 54. automation script to csv insertion

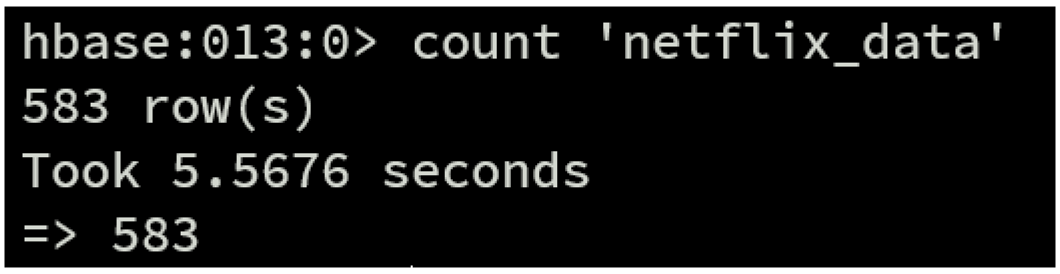
list and run py3 script to insert_data.py



```
hbase:003:0> list
TABLE
netflix_data
1 row(s)
Took 3.1579 seconds
=> ["netflix_data"]
hbase:004:0> exit
hadoop@DESKTOP-O5BDU29:/usr/local/HBase/bin$ cd
hadoop@DESKTOP-O5BDU29:~$ python3 insert_data.py
100 rows inserted.
200 rows inserted.
300 rows inserted.
400 rows inserted.
500 rows inserted.
Total 583 rows inserted.
```

Figure 55. list table and run py3 script

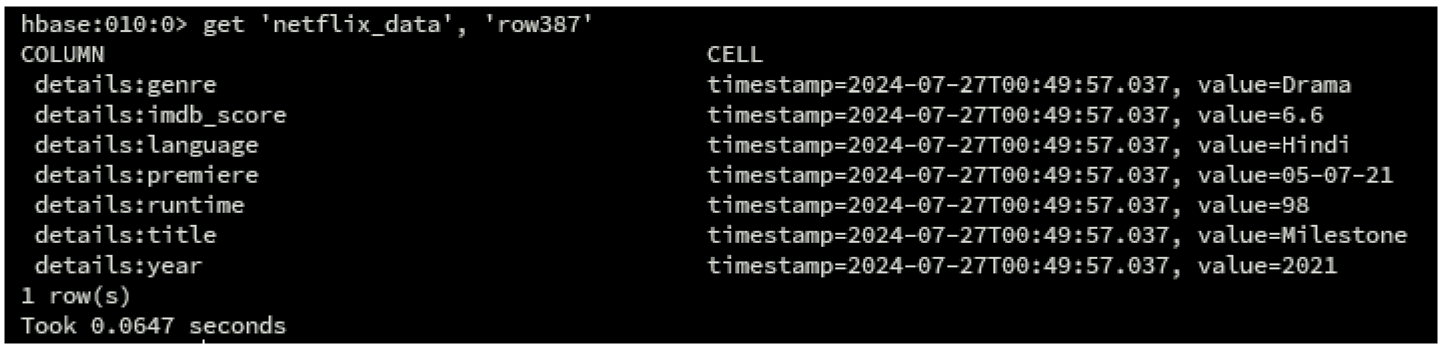
use count



```
hbase:013:0> count 'netflix_data'
583 row(s)
Took 5.5676 seconds
=> 583
```

Figure 56. count rows

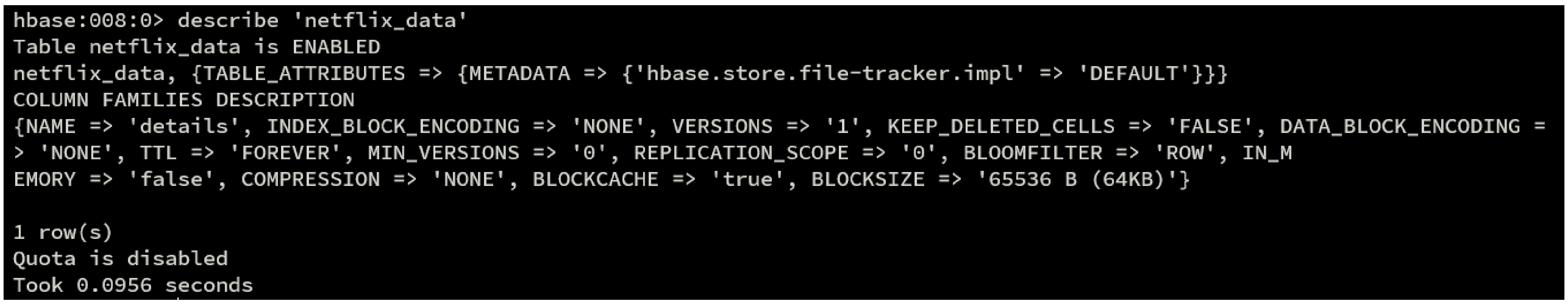
get specific row



```
hbase:010:0> get 'netflix_data', 'row387'
COLUMN                                                 CELL
 details:genre                                         timestamp=2024-07-27T00:49:57.037, value=Drama
 details:imdb_score                                    timestamp=2024-07-27T00:49:57.037, value=6.6
 details:language                                      timestamp=2024-07-27T00:49:57.037, value=Hindi
 details:premiere                                      timestamp=2024-07-27T00:49:57.037, value=05-07-21
 details:runtime                                       timestamp=2024-07-27T00:49:57.037, value=98
 details:title                                         timestamp=2024-07-27T00:49:57.037, value=Milestone
 details:year                                          timestamp=2024-07-27T00:49:57.037, value=2021
1 row(s)
Took 0.0647 seconds
```

Figure 57. get specific row

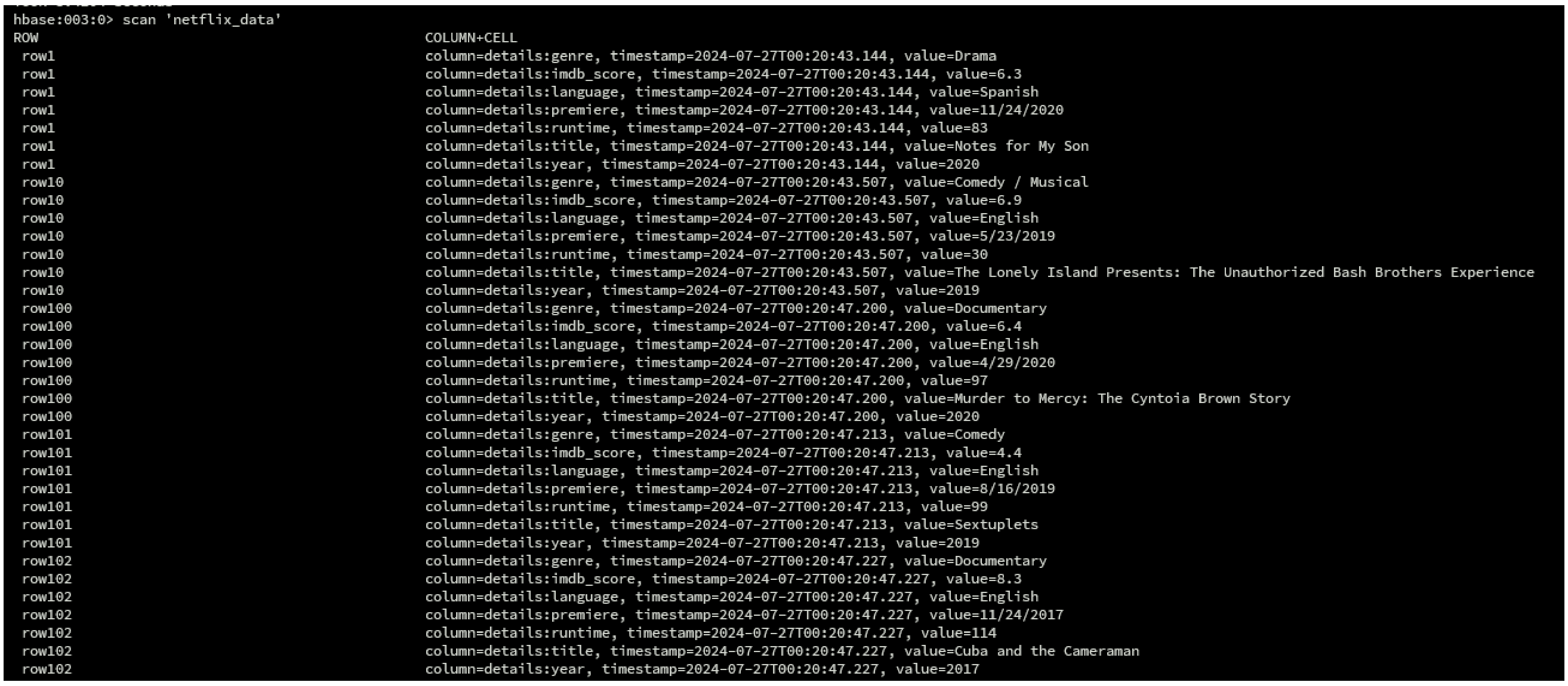
describe



```
hbase:008:0> describe 'netflix_data'
Table netflix_data is ENABLED
netflix_data, {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}
COLUMN FAMILIES DESCRIPTION
{NAME => 'details', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING =
> 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_M
EMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

1 row(s)
Quota is disabled
Took 0.0956 seconds
```

Figure 58. describe csv

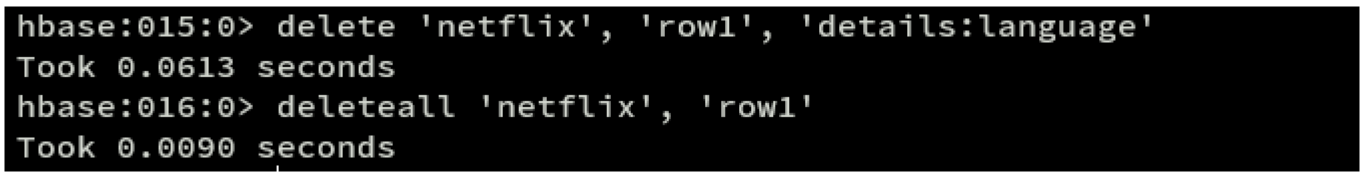
scan



```
hbase:003:0> scan 'netflix_data'
ROW                                                COLUMN+CELL
 row1                                              column=details:genre, timestamp=2024-07-27T00:20:43.144, value=Drama
 row1                                              column=details:imdb_score, timestamp=2024-07-27T00:20:43.144, value=6.3
 row1                                              column=details:language, timestamp=2024-07-27T00:20:43.144, value=Spanish
 row1                                              column=details:premiere, timestamp=2024-07-27T00:20:43.144, value=11/24/2020
 row1                                              column=details:runtime, timestamp=2024-07-27T00:20:43.144, value=83
 row1                                              column=details:title, timestamp=2024-07-27T00:20:43.144, value=Notes for My Son
 row1                                              column=details:year, timestamp=2024-07-27T00:20:43.144, value=2020
 row10                                             column=details:genre, timestamp=2024-07-27T00:20:43.507, value=Comedy / Musical
 row10                                             column=details:imdb_score, timestamp=2024-07-27T00:20:43.507, value=6.9
 row10                                             column=details:language, timestamp=2024-07-27T00:20:43.507, value=English
 row10                                             column=details:premiere, timestamp=2024-07-27T00:20:43.507, value=5/23/2019
 row10                                             column=details:runtime, timestamp=2024-07-27T00:20:43.507, value=30
 row10                                             column=details:title, timestamp=2024-07-27T00:20:43.507, value=The Lonely Island Presents: The Unauthorized Bash Brothers Experience
 row10                                             column=details:year, timestamp=2024-07-27T00:20:43.507, value=2019
 row100                                            column=details:genre, timestamp=2024-07-27T00:20:47.200, value=Documentary
 row100                                            column=details:imdb_score, timestamp=2024-07-27T00:20:47.200, value=6.4
 row100                                            column=details:language, timestamp=2024-07-27T00:20:47.200, value=English
 row100                                            column=details:premiere, timestamp=2024-07-27T00:20:47.200, value=4/29/2020
 row100                                            column=details:runtime, timestamp=2024-07-27T00:20:47.200, value=97
 row100                                            column=details:title, timestamp=2024-07-27T00:20:47.200, value=Murder to Mercy: The Cyntoia Brown Story
 row100                                            column=details:year, timestamp=2024-07-27T00:20:47.200, value=2020
 row101                                            column=details:genre, timestamp=2024-07-27T00:20:47.213, value=Comedy
 row101                                            column=details:imdb_score, timestamp=2024-07-27T00:20:47.213, value=4.4
 row101                                            column=details:language, timestamp=2024-07-27T00:20:47.213, value=English
 row101                                            column=details:premiere, timestamp=2024-07-27T00:20:47.213, value=8/16/2019
 row101                                            column=details:runtime, timestamp=2024-07-27T00:20:47.213, value=99
 row101                                            column=details:title, timestamp=2024-07-27T00:20:47.213, value=Sextuplets
 row101                                            column=details:year, timestamp=2024-07-27T00:20:47.213, value=2019
 row102                                            column=details:genre, timestamp=2024-07-27T00:20:47.227, value=Documentary
 row102                                            column=details:imdb_score, timestamp=2024-07-27T00:20:47.227, value=8.3
 row102                                            column=details:language, timestamp=2024-07-27T00:20:47.227, value=English
 row102                                            column=details:premiere, timestamp=2024-07-27T00:20:47.227, value=11/24/2017
 row102                                            column=details:runtime, timestamp=2024-07-27T00:20:47.227, value=114
 row102                                            column=details:title, timestamp=2024-07-27T00:20:47.227, value=Cuba and the Cameraman
 row102                                            column=details:year, timestamp=2024-07-27T00:20:47.227, value=2017
```

Figure 59. scan csv

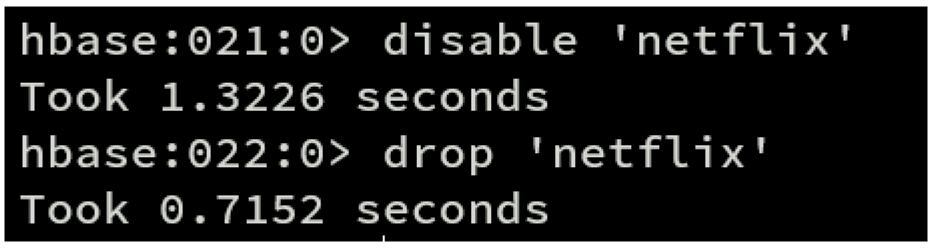
delete



```
hbase:015:0> delete 'netflix', 'row1', 'details:language'
Took 0.0613 seconds
hbase:016:0> deleteall 'netflix', 'row1'
Took 0.0090 seconds
```

Figure 60. delete row and whole csv

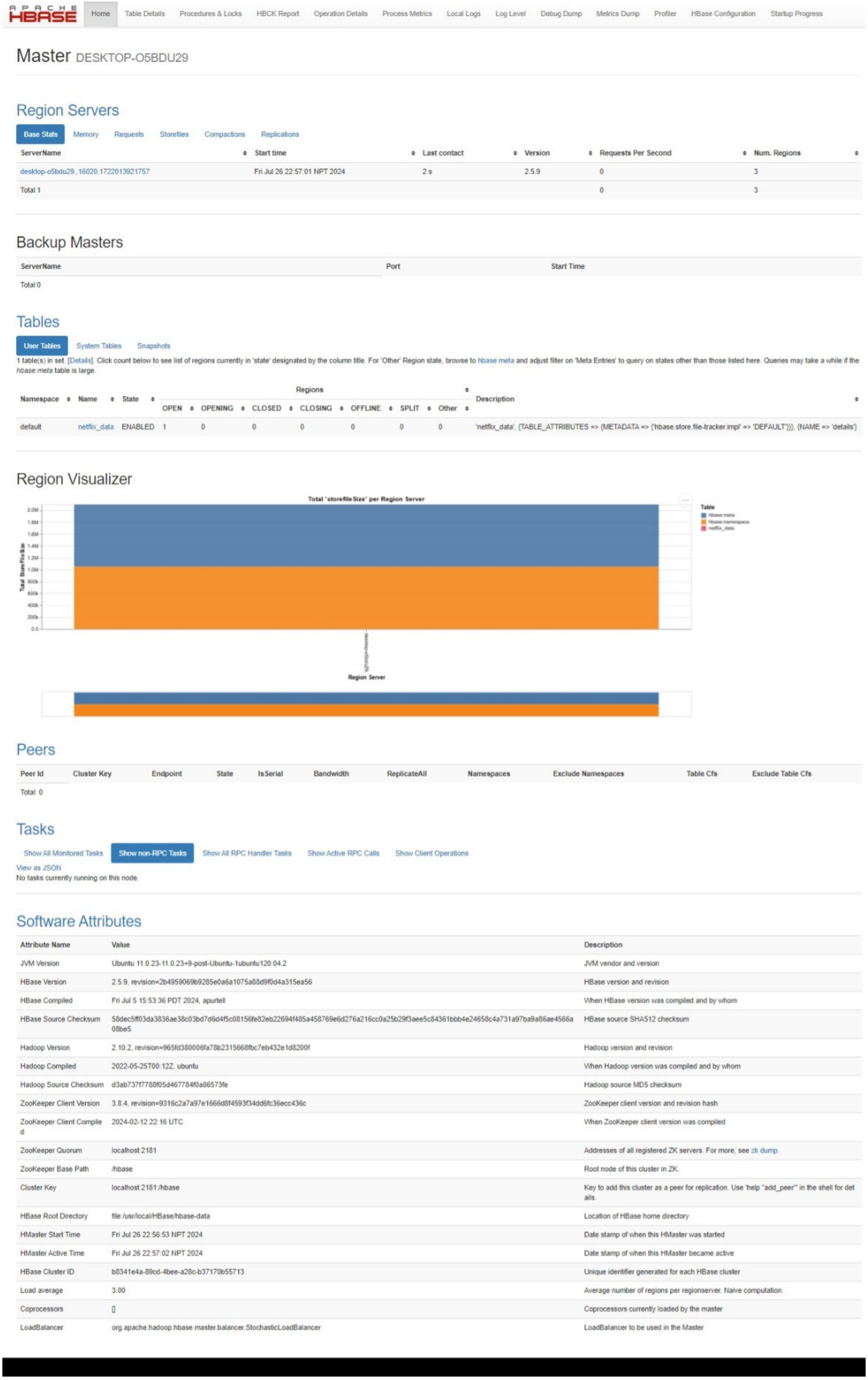
disable drop



```
hbase:021:0> disable 'netflix'
Took 1.3226 seconds
hbase:022:0> drop 'netflix'
Took 0.7152 seconds
```

Figure 61. disable and drop table

seeing localhost 16910 for hbase web interface



Master DESKTOP-O5BDU29

Region Servers

ServerName	Start time	Last contact	Version	Requests Per Second	Num. Regions
desktop-o5bdu29.,16020,1722013921757	Fri Jul 26 22:57:01 NPT 2024	2 s	2.5.9	0	3
Total:1				0	3

Backup Masters

ServerName	Port	Start Time
Total:0		

Tables

1 table(s) in set. [Details]. Click count below to see list of regions currently in 'state' designated by the column title. For 'Other' Region state, browse to hbase:meta and adjust filter on 'Meta Entries' to query on states other than those listed here. Queries may take a while if the *hbase:meta* table is large.

Namespace	Name	State	Regions							Description
			OPEN	OPENING	CLOSED	CLOSING	OFFLINE	SPLIT	Other	
default	netflix_data	ENABLED	1	0	0	0	0	0	0	'netflix_data', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'details'}

Region Visualizer

Peers

Peer Id	Cluster Key	Endpoint	State	IsSerial	Bandwidth	ReplicateAll	Namespaces	Exclude Namespaces	Table Cfs	Exclude Table Cfs
Total: 0										

Tasks

No tasks currently running on this node.

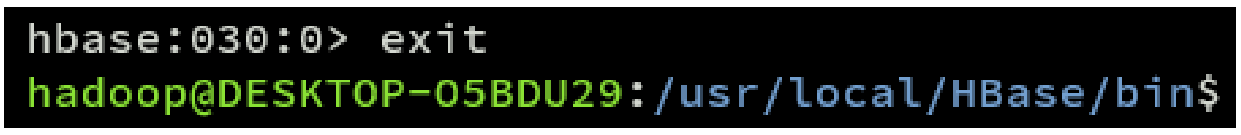
Software Attributes

Attribute Name	Value	Description
JVM Version	Ubuntu 11.0.23-11.0.23+9-post-Ubuntu-1ubuntu120.04.2	JVM vendor and version
HBase Version	2.5.9, revision=2b4959069b9285e0a6a1075a88d9f0d4a315ea56	HBase version and revision
HBase Compiled	Fri Jul 5 15:53:36 PDT 2024, apurtell	When HBase version was compiled and by whom
HBase Source Checksum	58dec5ff03da3836ae38c03bd7d6d4f5c08156fe82eb22694f485a458769e6d276a216cc0a25b29f3aee5c84361bbb4e24658c4a731a97ba9a86ae4566a 08be5	HBase source SHA512 checksum
Hadoop Version	2.10.2, revision=965fd380006fa78b2315668fbc7eb432e1d8200f	Hadoop version and revision
Hadoop Compiled	2022-05-25T00:12Z, ubuntu	When Hadoop version was compiled and by whom
Hadoop Source Checksum	d3ab737f7788f05d467784f0a86573fe	Hadoop source MD5 checksum
ZooKeeper Client Version	3.8.4, revision=9316c2a7a97e1666d8f4593f34dd6fc36ecc436c	ZooKeeper client version and revision hash
ZooKeeper Client Compile d	2024-02-12 22:16 UTC	When ZooKeeper client version was compiled
ZooKeeper Quorum	localhost:2181	Addresses of all registered ZK servers. For more, see zk dump.
ZooKeeper Base Path	/hbase	Root node of this cluster in ZK.
Cluster Key	localhost:2181:/hbase	Key to add this cluster as a peer for replication. Use 'help "add_peer"' in the shell for det ails.
HBase Root Directory	file:/usr/local/HBase/hbase-data	Location of HBase home directory
HMaster Start Time	Fri Jul 26 22:56:53 NPT 2024	Date stamp of when this HMaster was started
HMaster Active Time	Fri Jul 26 22:57:02 NPT 2024	Date stamp of when this HMaster became active
HBase Cluster ID	b8341e4a-89cd-4bee-a28c-b37170b55713	Unique identifier generated for each HBase cluster
Load average	3.00	Average number of regions per regionserver. Naive computation.
Coprocessors	[]	Coprocessors currently loaded by the master
LoadBalancer	org.apache.hadoop.hbase.master.balancer.StochasticLoadBalancer	LoadBalancer to be used in the Master

Figure 62. hbase webinterface

7.5. Wrap Up

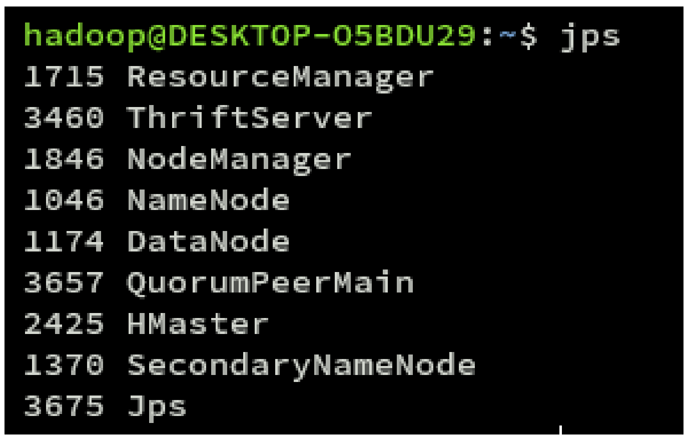
exit hbase



```
hbase:030:0> exit
hadoop@DESKTOP-05BDU29:/usr/local/HBase/bin$
```

Figure 63. exit hbase

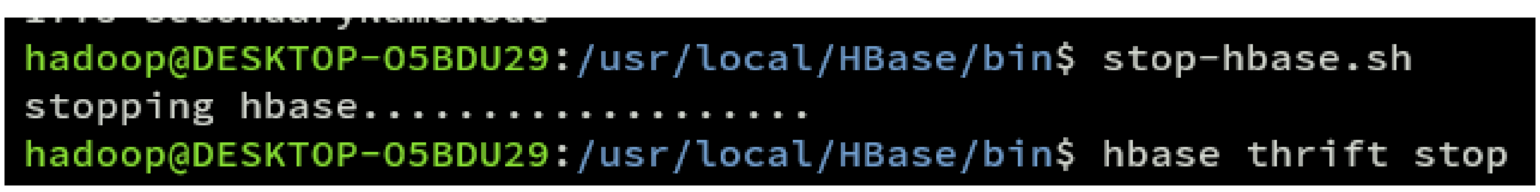
check jps



```
hadoop@DESKTOP-05BDU29:~$ jps
1715 ResourceManager
3460 ThriftServer
1846 NodeManager
1046 NameNode
1174 DataNode
3657 QuorumPeerMain
2425 HMaster
1370 SecondaryNameNode
3675 Jps
```

Figure 64. check running process

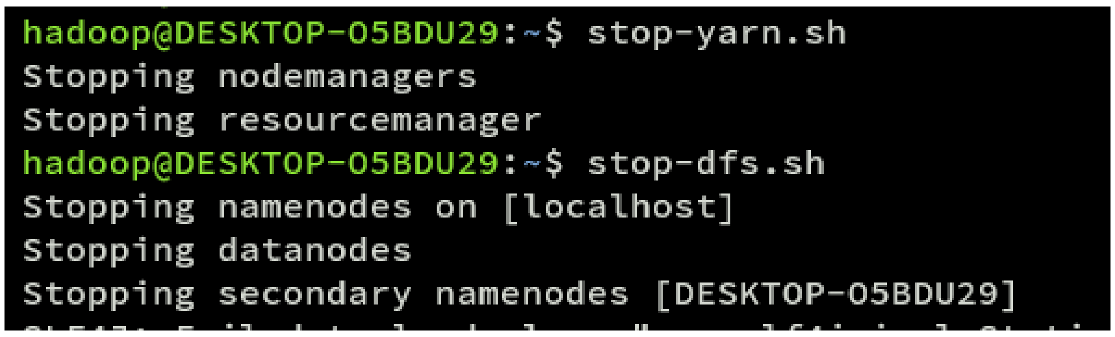
stop-hbase.sh and thrift



```
hadoop@DESKTOP-05BDU29:/usr/local/HBase/bin$ stop-hbase.sh
stopping hbase...................
hadoop@DESKTOP-05BDU29:/usr/local/HBase/bin$ hbase thrift stop
```

Figure 65. stop hbase and thrift

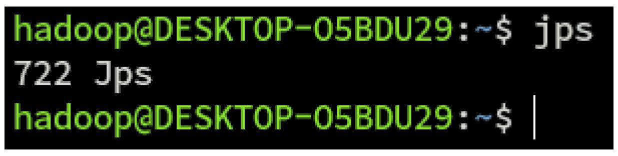
stop-yarn and dfs.sh



```
hadoop@DESKTOP-05BDU29:~$ stop-yarn.sh
Stopping nodemanagers
Stopping resourcemanager
hadoop@DESKTOP-05BDU29:~$ stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [DESKTOP-05BDU29]
```

Figure 66. stop dfs and yarn

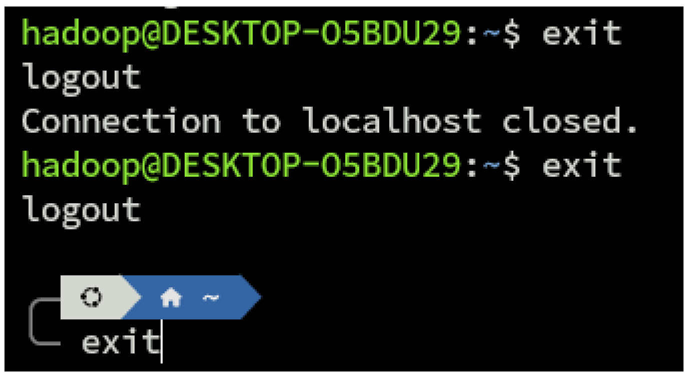
check jps



```
hadoop@DESKTOP-05BDU29:~$ jps
722 Jps
hadoop@DESKTOP-05BDU29:~$
```

Figure 67. check jps

exit localhost hadoop user and ubuntu terminal



```
hadoop@DESKTOP-05BDU29:~$ exit
logout
Connection to localhost closed.
hadoop@DESKTOP-05BDU29:~$ exit
logout

   ~
  exit
```

Figure 68. exit all

8. Analysis Report

Analysis of the Netflix dataset informs about the content of the streaming platform. Parameters like genre, premiere date, runtime, IMDb score, language, and release year give valuable insights into trends of viewing and content preference. We could process and analyze this large dataset with the help of Hadoop ecosystem tools.

Hive Analysis: We used Hive queries for high-rated movies, an average IMDb rating in a category, recent releases, and the length of movies categorized. This helped understand the performance of content and audience preferences.

Pig Analysis: Filtering of movies with release year and IMDb score was done with the help of Pig scripts, followed by grouping data by genre and creating derived columns of runtime categories. This would allow for detailed data transformation and aggregation

HBase Analysis: HBase was used for storing and retrieving data; running several queries on it to show its efficiency—like creating tables, inserting data, and scanning to retrieve rows.

8.1. Recommendation/r

Improvement in Data Integration:

- Focus on bettering integration processes across different Hadoop ecosystem components, such as Hive, Pig, and HBase, to improve data workflows.
- Put in place automated data validation checks at the time of ingestion to ensure data quality.

Advanced Analytical Techniques:

- Run machine learning models within Hive and Pig queries in order to predict content trends in popularity and help with improvement of recommendation algorithms.
- Use sentiment analysis based on reviews and ratings given by users to fine-tune content recommendations.

Scalability Enhancements:

- Leverage cloud-based Hadoop to scale resources dynamically relative to the size of data or processing.
- Keep updating cluster configurations at regular intervals to achieve optimal performance and cost efficiency.

User Personalization:

- Use user behavior viewing patterns and preferences to generate more relevant content recommendations.
- Implement real-time analytics that respond in milliseconds with suggestions of relevant, personalized content.

8.2. Research Opportunities

Content Popularity Prediction:

- Develop predictive models to use past data and forecast the popularity of new releases.
- Conduct research into genre-based viewing patterns to understand the content preferences across different demographics.

Enhanced Recommendation System:

- Research hybrid recommender systems that can combine both collaborative and content-based filtering.
- Steadily enhance the recommendation accuracy with regard to investigating the effects of temporal dynamics on user preferences.

Data Processing Optimizations:

- It is worth investing in research into new data processing frameworks that would outclass the current Hadoop-based solutions in performance and efficiency.
- Study how quantum computing can be used in big data analytics for the better handling of huge datasets.

Sociocultural Impact Analysis:

- The representation of cultures in Netflix programming with respect to its impact on viewership.
- The role of local content in increasing the number of subscribers should be explored for different regions.

9. Conclusion

Analysis of the Netflix dataset using Hadoop, Hive, Pig, and HBase has returned a number of insights related to content trends and user preferences. Netflix can further improve upon its recommendation system and content strategy by implementing further improvements in data integration, applying advanced analytics, and scaling up the underlying infrastructure. Further predictive modeling research into recommendation systems and data processing optimization will only continue the fine-tuning of these processes, keeping Netflix at the top in the streaming industry.

9.1. Evidence.

The Hadoop ecosystem components selected, consisting of Hadoop Distributed File System, Hive, Pig, and HBase, were practical in handling the large Netflix dataset and running its analysis. This is because these tools have robust abilities in distributed processing, SQL-like queries, complex data transformations, and efficient storing and retrieving of data. The successful running of various queries and scripts illustrates their capability to provide meaningful insights and support data-driven decision-making.

Queries and Results:

- Hive: Created tables, loaded data, filtered high-rated movies, grouped by genre for average IMDb score, and categorized movies by runtime.
- Pig: Loaded dataset, filtered recent and high-rated movies, grouped by genre, and derived length categories.
- HBase: Created and populated tables, performed scans, and demonstrated quick data retrieval.

10. Reference

Amatriain, X. (2013). Big & personal: data and models behind Netflix recommendations. *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining Algorithms, Systems, Programming Models and Applications - BigMine '13*. https://doi.org/10.1145/2501221.2501222

Amazon. (2022). *AWS Innovator: Netflix | Case Studies, Videos and Customer Stories*. Amazon Web Services, Inc. https://aws.amazon.com/solutions/case-studies/innovators/netflix/

Bag Operations - Guide - Apache DataFu Pig. (n.d.). Datafu.apache.org. https://datafu.apache.org/docs/datafu/guide/bag-operations.html

Chauhan, Vaishali. (2016). "HIVE, PIG & HBASE PERFORMANCE EVALUATION FOR DATA PROCESSING APPLICATIONS".

Fouladirad, Maryam & Neal, John & Vilaplana Ituarte, Jorge & Alexander, Joshua & Ghareeb, Ahmad. (2019). Entertaining Data: Business Analytics and Netflix. Computer Science and Information Systems. 10. 13-22.

Hadoop Ecosystem: MapReduce, YARN, Hive, Pig, Spark, Oozie, Zookeeper, Mahout, and Kube2Hadoop - Bizety: Research & Consulting. (n.d.). https://www.bizety.com/2020/06/20/hadoop-ecosystem-mapreduce-yarn-hive-pig-spark-oozie-zookeeper-mahout-and-kube2hadoop/

Maddodi, S. (2019, October 21). *Netflix Bigdata Analytics - The Emergence of Data Driven Recommendation*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3473148

Mahmood, Z. (2016). Data Science and Big Data Computing. In *Springer eBooks*. Springer Nature. https://doi.org/10.1007/978-3-319-31861-5

N. Maheswari, & M. Sivagami. (2016). Large-Scale data analytics tools: Apache hive, pig, and hbase. *Springer EBooks*, 191–220. https://doi.org/10.1007/978-3-319-31861-5_9

Netflix. (2018). *Netflix Research*. Netflix.com. https://research.netflix.com/

Netflix. (2019). *Netflix | Open Connect*. Netflix.com. https://openconnect.netflix.com/en/

Netflix Technology Blog. (n.d.). Medium. https://netflixtechblog.medium.com/

Shivalkar, R. (2024, April 4). *Netflix Architecture | A Look Into Its System Architecture*. ClickIT. https://www.clickittech.com/application-architecture/netflix-architecture/?utm_source=medium&utm_medium=referral

Storage using Amazon S3 | Netflix Video | AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/solutions/case-studies/netflix-storage-reinvent22/

Vergilio, T., & Ramachandran, M. (2018). Non-functional Requirements for Real World Big Data Systems - An Investigation of Big Data Architectures at Facebook, Twitter and Netflix. *Proceedings of the 13th International Conference on Software Technologies*. https://doi.org/10.5220/0006825408330840