Methodology

Document Processing Pipeline

The implementation is designed to process documents using a multi-stage MapReduce pipeline, primarily focusing on text data extraction, organization, and aggregation to facilitate document-centric analysis. Initially, the first mapper extracts key metadata and terms from documents by parsing document IDs and titles from filenames. It tokenizes the text using a regex pattern to delineate word boundaries and filters out single-character words to minimize noise. The output from this stage consists of tuples of the form (term, docID, title, count) for each word in the document.

The subsequent intermediary mapping stage reorganizes these tuples to create records centered around each document, altering them to (docID, term, title, count). This restructuring lays the groundwork for subsequent aggregation during the reduction phases. The first reducer focuses on aggregating term statistics by compiling occurrences of the same term within a document to maintain a record of term frequencies. The final reducer then generates extensive document statistics by calculating each document's length, documenting term frequencies, and tracking document frequency for each term, necessary for inverse document frequency (IDF) calculations. The final outputs from the reducer include document metadata records, term frequency records, and term document count records, providing a comprehensive dataset for further analysis.

Data Storage Approach

The implementation leverages Cassandra as the storage backend for the search index, optimized for scalability and efficient read operations. This setup involves storing document metadata (such as ID, title, and length) in a documents table, term frequencies in a term_frequencies table indexed by term and document ID, and document frequencies in a term_document_count table. This schema is designed to facilitate efficient retrieval of necessary statistics for BM25 calculations during query processing, supporting rapid and scalable access to the stored data.

Query Processing Implementation

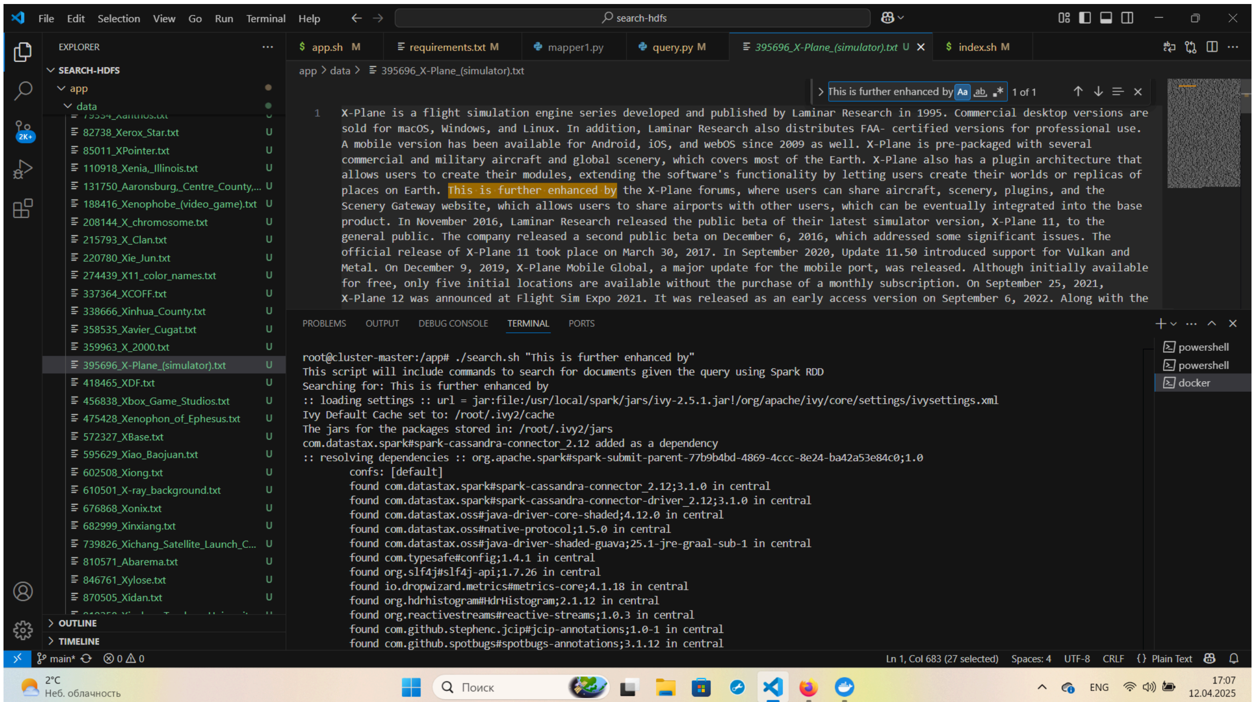
For query processing, Apache Spark is employed to enable distributed processing. The process begins with query analysis, where user queries are parsed using the same tokenization method employed in document processing and converted to lowercase to ensure consistent matching. The retrieval model implements the BM25 ranking function with parameters k1=1.2 and b=0.75, which are standard starting points for BM25; these parameters control term frequency saturation and document length normalization. Using Spark, the score calculation is distributed across worker nodes, with each node independently calculating the BM25 score for documents by utilizing term frequency from the term_frequencies table, document frequency from the term_document_count table, and document length statistics from the documents table. The standard BM25 formula is used to compute the scores. Finally, results are filtered to include only documents with positive scores, sorted in descending order to prioritize relevance, and the top 10 documents, along with their titles and scores, are returned to the user.

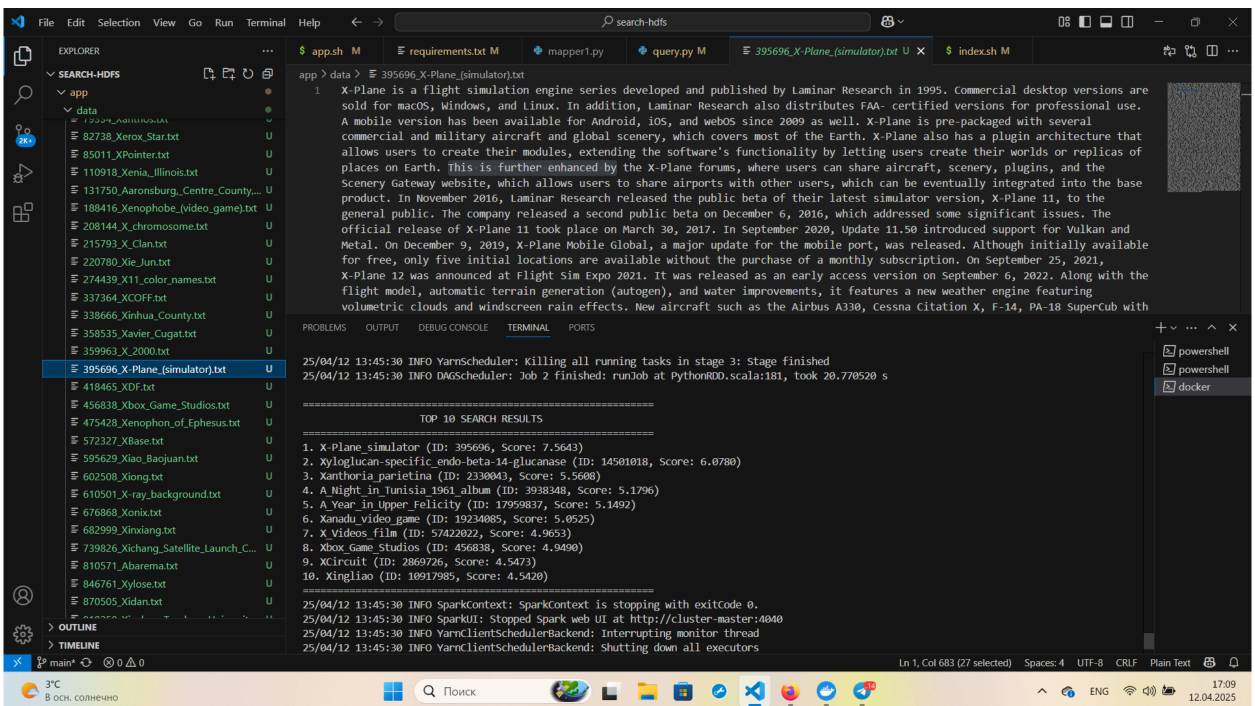
Demo

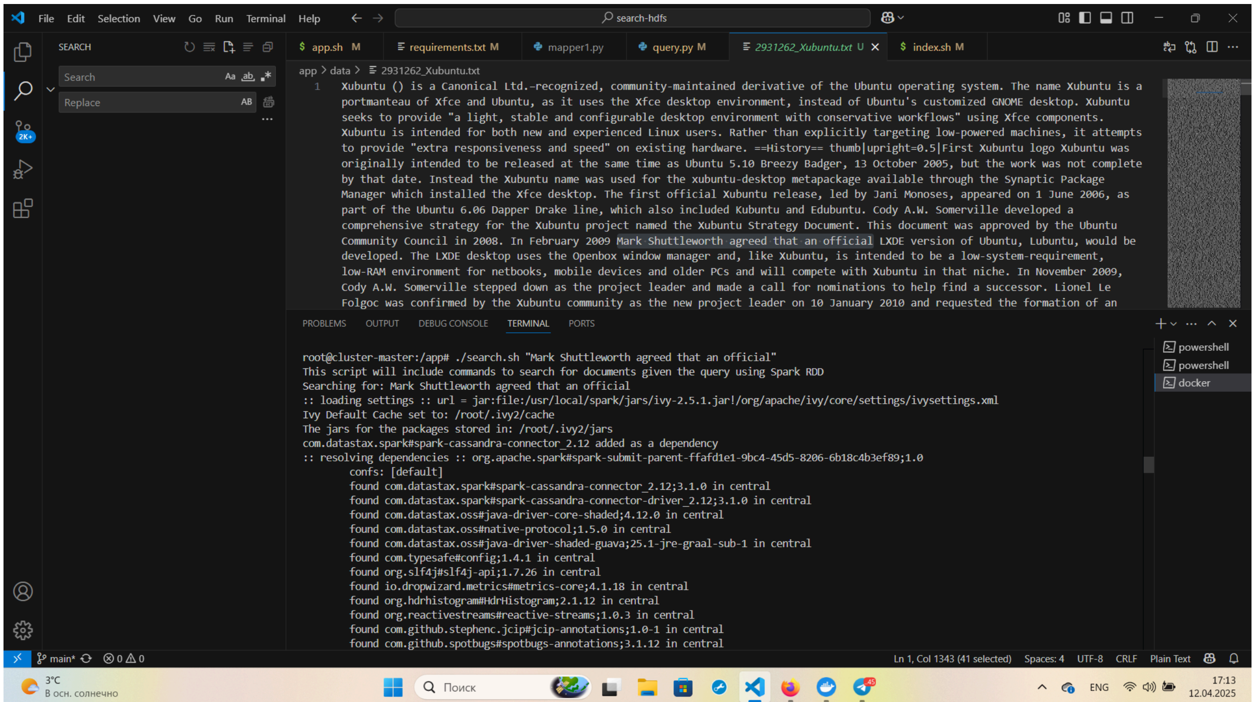
https://github.com/EninDmitriy96/search-hdfs

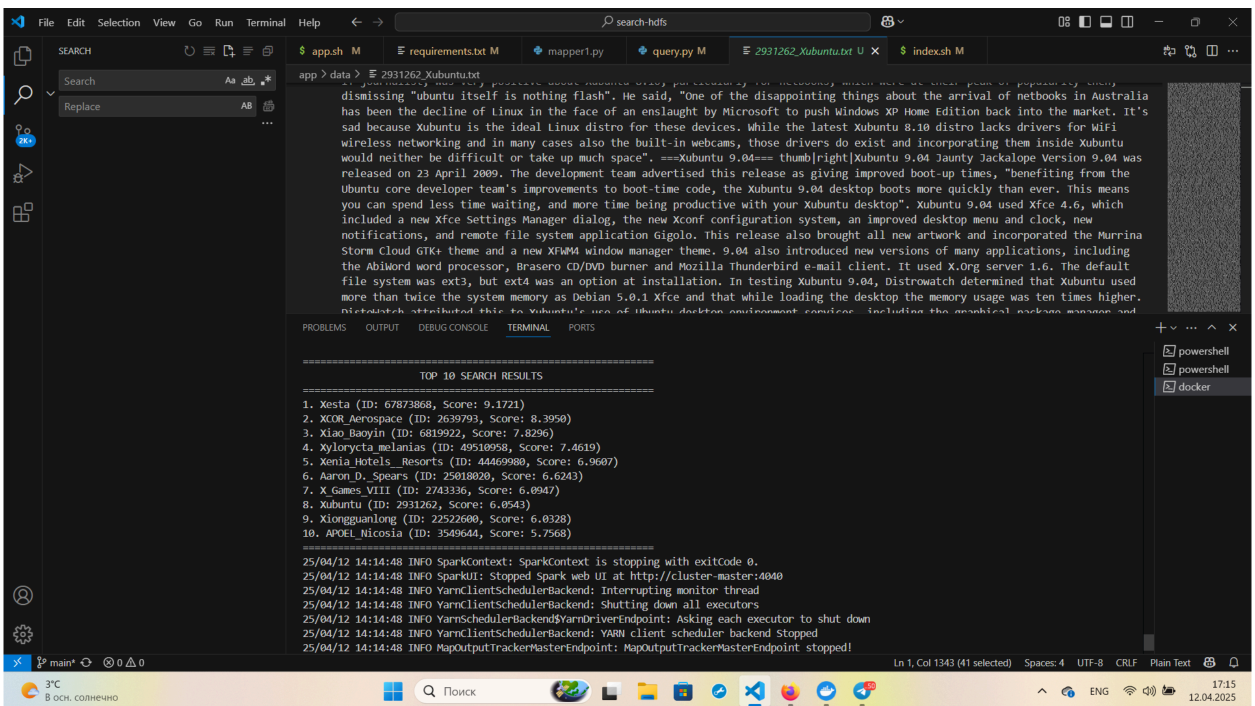
The implementation expects to be run inside the docker container. First, one has to build it using `docker compose up -d`. After that, enter the container with `docker exec -it cluster-master bash -c`. Within the container, application may be run via `./app.sh`. This will start the services up, upload data from `/data` folder to HDFS (expected format `<doc_id>_<doc_title>.txt`), run the index procedure and upload to cassandra. In the end, the app.sh will run a sample search. Further search may be performed via `./search.sh "search query"`, top results will be displayed (among the logs, can be found closer to the end of the output).

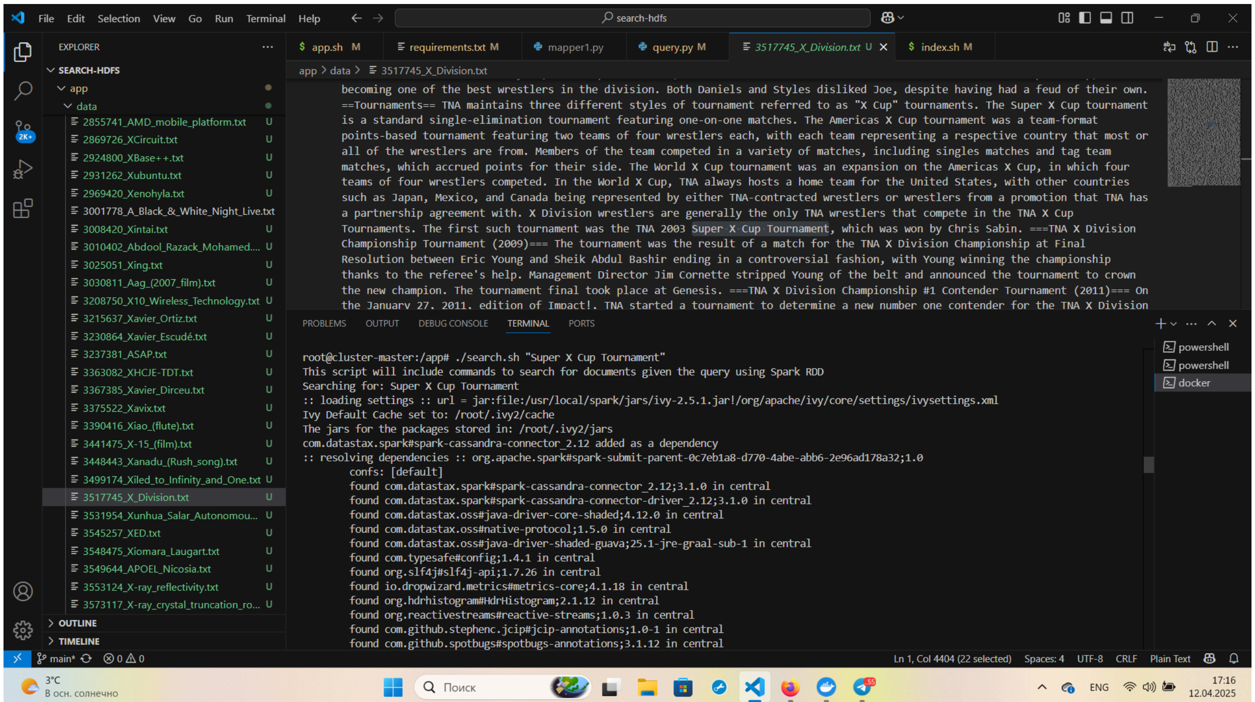
Document size can significantly impact search relevance, even when searching for exact phrases. This counterintuitive behavior occurs because most modern search engines employ length normalization in their ranking algorithms (like BM25), which penalizes term matches in longer documents to prevent them from dominating results solely due to their higher term frequency potential. When a document is substantially longer, your exact phrase represents a smaller percentage of the overall content, resulting in a lower normalized relevance score despite being a perfect match. Additionally, longer documents may have more competing matches for partial terms from query, and the exact phrase might appear in a less prominent section (like deep in the body text rather than in headings or introductory paragraphs). Search engines also consider term proximity, so if your phrase appears once in a short document versus once in a lengthy document with thousands of other words, the shorter document might rank higher because the matched terms represent a more significant proportion of its content. See the example with search on phrase from Xubuntu file. The screenshots are attached below.

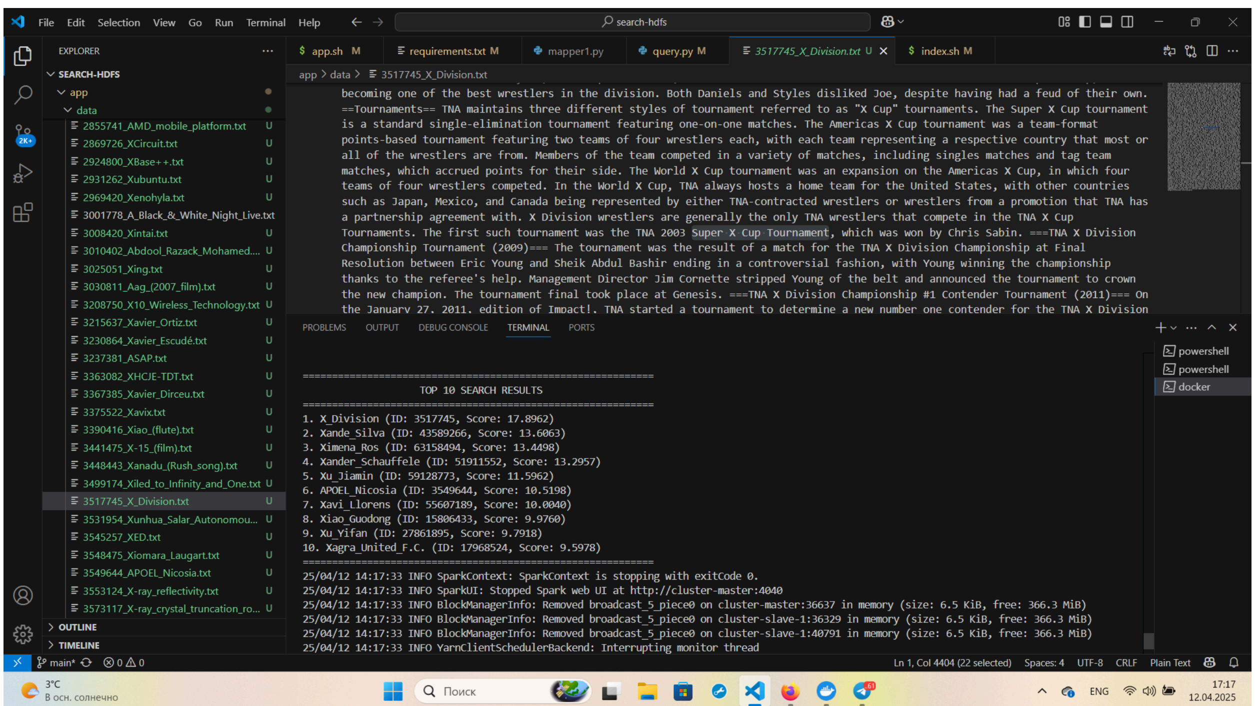

File Edit Selection View Go Run Terminal Help
search-hdfs
EXPLORER
SEARCH-HDFS
app
data
82738_Xerox_Star.txt
85011_XPointer.txt
110918_Xenia,_Illinois.txt
131750_Aaronsburg,_Centre_County,...
188416_Xenophobe_(video_game).txt
208144_X_chromosome.txt
215793_X_Clan.txt
220780_Xie_Jun.txt
274439_X11_color_names.txt
337364_XCOFF.txt
338666_Xinhua_County.txt
358535_Xavier_Cugat.txt
359963_X_2000.txt
395696_X-Plane_(simulator).txt
418465_XDF.txt
456838_Xbox_Game_Studios.txt
475428_Xenophon_of_Ephesus.txt
572327_XBase.txt
595629_Xiao_Baojuan.txt
602508_Xiong.txt
610501_X-ray_background.txt
676868_Xonix.txt
682999_Xinxiang.txt
739826_Xichang_Satellite_Launch_C...
810571_Abarema.txt
846761_Xylose.txt
870505_Xidan.txt
OUTLINE
TIMELINE
app.sh M
requirements.txt M
mapper1.py
query.py M
395696_X-Plane_(simulator).txt U
index.sh M
app > data > 395696_X-Plane_(simulator).txt
This is further enhanced by
1 of 1
X-Plane is a flight simulation engine series developed and published by Laminar Research in 1995. Commercial desktop versions are sold for macOS, Windows, and Linux. In addition, Laminar Research also distributes FAA- certified versions for professional use. A mobile version has been available for Android, iOS, and webOS since 2009 as well. X-Plane is pre-packaged with several commercial and military aircraft and global scenery, which covers most of the Earth. X-Plane also has a plugin architecture that allows users to create their modules, extending the software's functionality by letting users create their worlds or replicas of places on Earth. This is further enhanced by the X-Plane forums, where users can share aircraft, scenery, plugins, and the Scenery Gateway website, which allows users to share airports with other users, which can be eventually integrated into the base product. In November 2016, Laminar Research released the public beta of their latest simulator version, X-Plane 11, to the general public. The company released a second public beta on December 6, 2016, which addressed some significant issues. The official release of X-Plane 11 took place on March 30, 2017. In September 2020, Update 11.50 introduced support for Vulkan and Metal. On December 9, 2019, X-Plane Mobile Global, a major update for the mobile port, was released. Although initially available for free, only five initial locations are available without the purchase of a monthly subscription. On September 25, 2021, X-Plane 12 was announced at Flight Sim Expo 2021. It was released as an early access version on September 6, 2022. Along with the
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
root@cluster-master:/app# ./search.sh "This is further enhanced by"
This script will include commands to search for documents given the query using Spark RDD
Searching for: This is further enhanced by
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-77b9b4bd-4869-4ccc-8e24-ba42a53e84c0;1.0
confs: [default]
found com.datastax.spark#spark-cassandra-connector_2.12;3.1.0 in central
found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.1.0 in central
found com.datastax.oss#java-driver-core-shaded;4.12.0 in central
found com.datastax.oss#native-protocol;1.5.0 in central
found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
found com.typesafe#config;1.4.1 in central
found org.slf4j#slf4j-api;1.7.26 in central
found io.dropwizard.metrics#metrics-core;4.1.18 in central
found org.hdrhistogram#HdrHistogram;2.1.12 in central
found org.reactivestreams#reactive-streams;1.0.3 in central
found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
powershell
powershell
docker
main* 0 0
Ln 1, Col 683 (27 selected) Spaces: 4 UTF-8 CRLF Plain Text
2°C
Неб. облачность
Поиск
ENG
17:07
12.04.2025

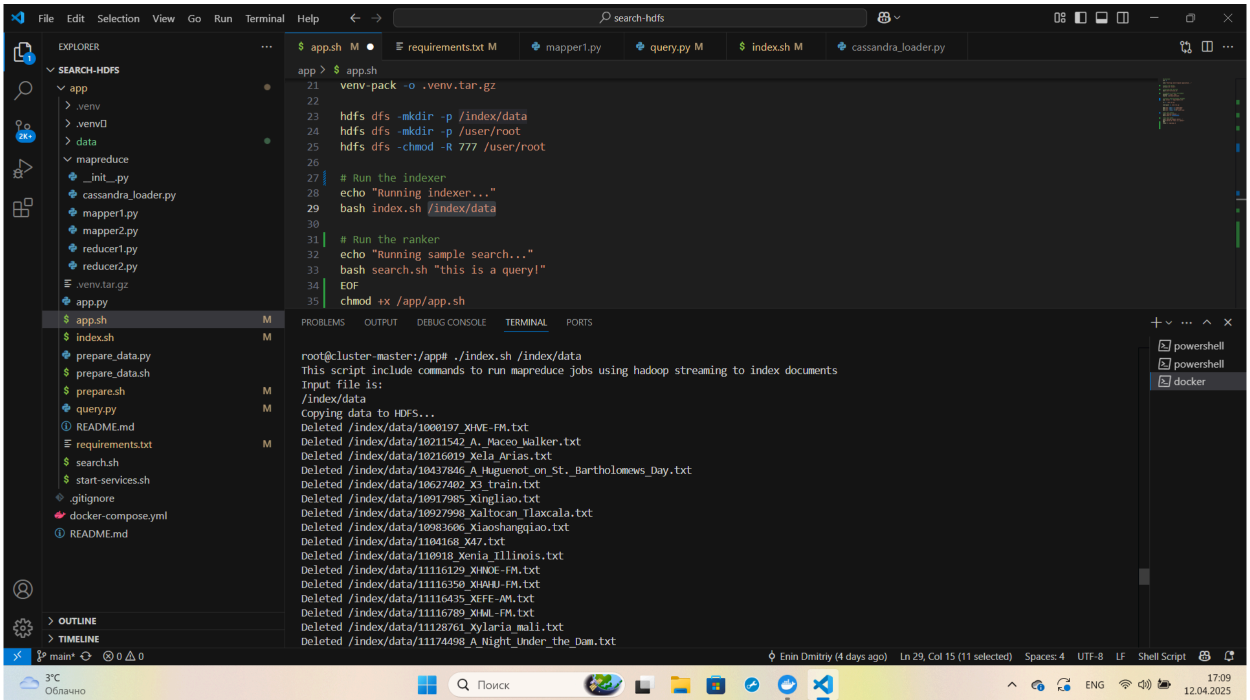

File Edit Selection View Go Run Terminal Help
search-hdfs
EXPLORER
SEARCH-HDFS
app
data
app > data > 395696_X-Plane_(simulator).txt
X-Plane is a flight simulation engine series developed and published by Laminar Research in 1995. Commercial desktop versions are sold for macOS, Windows, and Linux. In addition, Laminar Research also distributes FAA- certified versions for professional use. A mobile version has been available for Android, iOS, and webOS since 2009 as well. X-Plane is pre-packaged with several commercial and military aircraft and global scenery, which covers most of the Earth. X-Plane also has a plugin architecture that allows users to create their modules, extending the software's functionality by letting users create their worlds or replicas of places on Earth. This is further enhanced by the X-Plane forums, where users can share aircraft, scenery, plugins, and the Scenery Gateway website, which allows users to share airports with other users, which can be eventually integrated into the base product. In November 2016, Laminar Research released the public beta of their latest simulator version, X-Plane 11, to the general public. The company released a second public beta on December 6, 2016, which addressed some significant issues. The official release of X-Plane 11 took place on March 30, 2017. In September 2020, Update 11.50 introduced support for Vulkan and Metal. On December 9, 2019, X-Plane Mobile Global, a major update for the mobile port, was released. Although initially available for free, only five initial locations are available without the purchase of a monthly subscription. On September 25, 2021, X-Plane 12 was announced at Flight Sim Expo 2021. It was released as an early access version on September 6, 2022. Along with the flight model, automatic terrain generation (autogen), and water improvements, it features a new weather engine featuring volumetric clouds and windscreen rain effects. New aircraft such as the Airbus A330, Cessna Citation X, F-14, PA-18 SuperCub with
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
25/04/12 13:45:30 INFO YarnScheduler: Killing all running tasks in stage 3: Stage finished
25/04/12 13:45:30 INFO DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:181, took 20.770520 s
==
TOP 10 SEARCH RESULTS
==
1. X-Plane_simulator (ID: 395696, Score: 7.5643)
2. Xyloglucan-specific_endo-beta-14-glucanase (ID: 14501018, Score: 6.0780)
3. Xanthoria_parietina (ID: 2330043, Score: 5.5608)
4. A_Night_in_Tunisia_1961_album (ID: 3938348, Score: 5.1796)
5. A_Year_in_Upper_Felicity (ID: 17959837, Score: 5.1492)
6. Xanadu_video_game (ID: 19234085, Score: 5.0525)
7. X_Videos_film (ID: 57422022, Score: 4.9653)
8. Xbox_Game_Studios (ID: 456838, Score: 4.9490)
9. XCircuit (ID: 2869726, Score: 4.5473)
10. Xingliao (ID: 10917985, Score: 4.5420)
==
25/04/12 13:45:30 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/12 13:45:30 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/12 13:45:30 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/12 13:45:30 INFO YarnClientSchedulerBackend: Shutting down all executors
powershell
powershell
docker
Ln 1, Col 683 (27 selected) Spaces: 4 UTF-8 CRLF Plain Text
3°C
В осн. солнечно
Поиск
ENG
17:09
12.04.2025


File Edit Selection View Go Run Terminal Help
search-hdfs
SEARCH
app.sh M
requirements.txt M
mapper1.py
query.py M
2931262_Xubuntu.txt U
index.sh M
app > data > 2931262_Xubuntu.txt
Xubuntu () is a Canonical Ltd.–recognized, community-maintained derivative of the Ubuntu operating system. The name Xubuntu is a portmanteau of Xfce and Ubuntu, as it uses the Xfce desktop environment, instead of Ubuntu's customized GNOME desktop. Xubuntu seeks to provide "a light, stable and configurable desktop environment with conservative workflows" using Xfce components. Xubuntu is intended for both new and experienced Linux users. Rather than explicitly targeting low-powered machines, it attempts to provide "extra responsiveness and speed" on existing hardware. ==History== thumb|upright=0.5|First Xubuntu logo Xubuntu was originally intended to be released at the same time as Ubuntu 5.10 Breezy Badger, 13 October 2005, but the work was not complete by that date. Instead the Xubuntu name was used for the xubuntu-desktop metapackage available through the Synaptic Package Manager which installed the Xfce desktop. The first official Xubuntu release, led by Jani Monoses, appeared on 1 June 2006, as part of the Ubuntu 6.06 Dapper Drake line, which also included Kubuntu and Edubuntu. Cody A.W. Somerville developed a comprehensive strategy for the Xubuntu project named the Xubuntu Strategy Document. This document was approved by the Ubuntu Community Council in 2008. In February 2009 Mark Shuttleworth agreed that an official LXDE version of Ubuntu, Lubuntu, would be developed. The LXDE desktop uses the Openbox window manager and, like Xubuntu, is intended to be a low-system-requirement, low-RAM environment for netbooks, mobile devices and older PCs and will compete with Xubuntu in that niche. In November 2009, Cody A.W. Somerville stepped down as the project leader and made a call for nominations to help find a successor. Lionel Le Folgoc was confirmed by the Xubuntu community as the new project leader on 10 January 2010 and requested the formation of an
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
root@cluster-master:/app# ./search.sh "Mark Shuttleworth agreed that an official"
This script will include commands to search for documents given the query using Spark RDD
Searching for: Mark Shuttleworth agreed that an official
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-ffafd1e1-9bc4-45d5-8206-6b18c4b3ef89;1.0
confs: [default]
found com.datastax.spark#spark-cassandra-connector_2.12;3.1.0 in central
found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.1.0 in central
found com.datastax.oss#java-driver-core-shaded;4.12.0 in central
found com.datastax.oss#native-protocol;1.5.0 in central
found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
found com.typesafe#config;1.4.1 in central
found org.slf4j#slf4j-api;1.7.26 in central
found io.dropwizard.metrics#metrics-core;4.1.18 in central
found org.hdrhistogram#HdrHistogram;2.1.12 in central
found org.reactivestreams#reactive-streams;1.0.3 in central
found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
powershell
powershell
docker
main* 0 0
Ln 1, Col 1343 (41 selected) Spaces: 4 UTF-8 CRLF Plain Text
3°C
В осн. солнечно
Поиск
ENG
17:13
12.04.2025

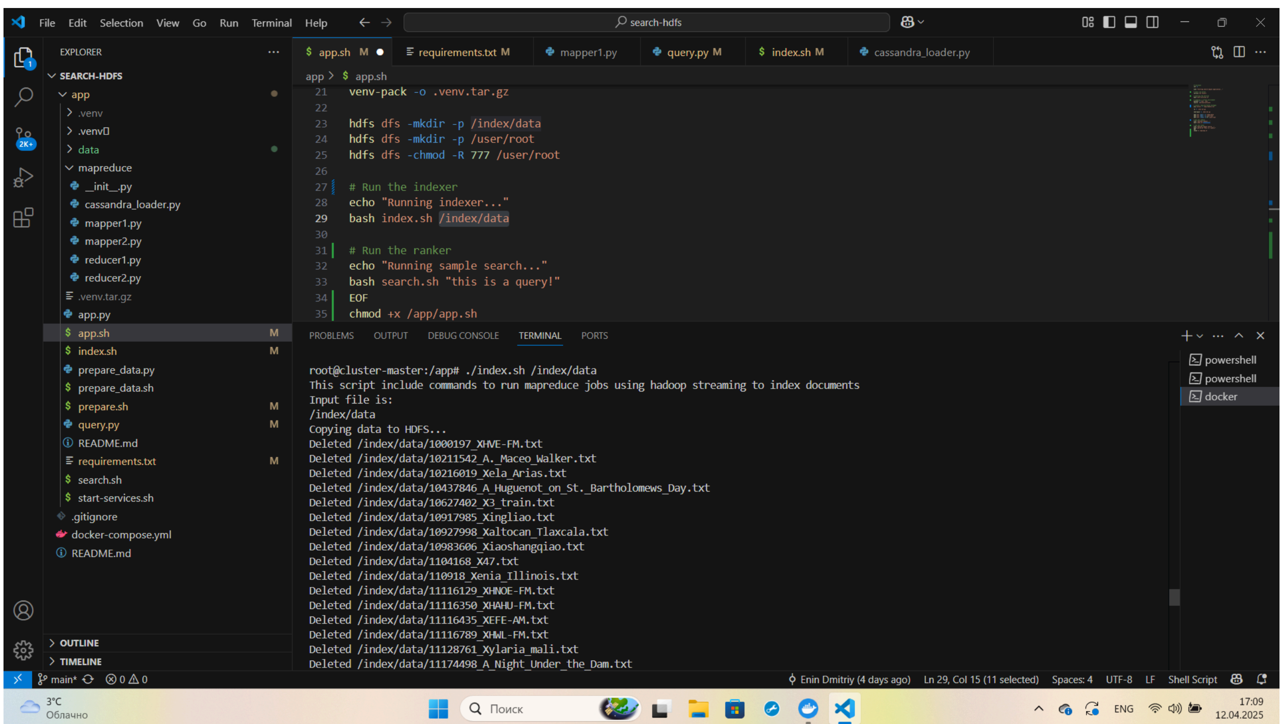

File Edit Selection View Go Run Terminal Help
search-hdfs
SEARCH
app.sh M
requirements.txt M
mapper1.py
query.py M
2931262_Xubuntu.txt U
index.sh M
app > data > 2931262_Xubuntu.txt
dismissing "ubuntu itself is nothing flash". He said, "One of the disappointing things about the arrival of netbooks in Australia has been the decline of Linux in the face of an enslaught by Microsoft to push Windows XP Home Edition back into the market. It's sad because Xubuntu is the ideal Linux distro for these devices. While the latest Xubuntu 8.10 distro lacks drivers for WiFi wireless networking and in many cases also the built-in webcams, those drivers do exist and incorporating them inside Xubuntu would neither be difficult or take up much space". ===Xubuntu 9.04=== thumb|right|Xubuntu 9.04 Jaunty Jackalope Version 9.04 was released on 23 April 2009. The development team advertised this release as giving improved boot-up times, "benefiting from the Ubuntu core development team's improvements to boot-time code, the Xubuntu 9.04 desktop boots more quickly than ever. This means you can spend less time waiting, and more time being productive with your Xubuntu desktop". Xubuntu 9.04 used Xfce 4.6, which included a new Xfce Settings Manager dialog, the new Xconf configuration system, an improved desktop menu and clock, new notifications, and remote file system application Gigolo. This release also brought all new artwork and incorporated the Murrina Storm Cloud GTK+ theme and a new XFWM4 window manager theme. 9.04 also introduced new versions of many applications, including the AbiWord word processor, Brasero CD/DVD burner and Mozilla Thunderbird e-mail client. It used X.Org server 1.6. The default file system was ext3, but ext4 was an option at installation. In testing Xubuntu 9.04, Distrowatch determined that Xubuntu used more than twice the system memory as Debian 5.0.1 Xfce and that while loading the desktop the memory usage was ten times higher.
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
==
TOP 10 SEARCH RESULTS
==
1. Xesta (ID: 67873868, Score: 9.1721)
2. XCOR_Aerospace (ID: 2639793, Score: 8.3950)
3. Xiao_Baoyin (ID: 6819922, Score: 7.8296)
4. Xylorycta_melanias (ID: 49510958, Score: 7.4619)
5. Xenia_Hotels__Resorts (ID: 44469980, Score: 6.9607)
6. Aaron_D._Spears (ID: 25018020, Score: 6.6243)
7. X_Games_VIII (ID: 2743336, Score: 6.0947)
8. Xubuntu (ID: 2931262, Score: 6.0543)
9. Xiongguanlong (ID: 22522600, Score: 6.0328)
10. APOEL_Nicosia (ID: 3549644, Score: 5.7568)
==
25/04/12 14:14:48 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/12 14:14:48 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/12 14:14:48 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/12 14:14:48 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/12 14:14:48 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/12 14:14:48 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/12 14:14:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
powershell
powershell
docker
main* 0 0
Ln 1, Col 1343 (41 selected) Spaces: 4 UTF-8 CRLF Plain Text
3°C
В осн. солнечно
Поиск
ENG
17:15
12.04.2025

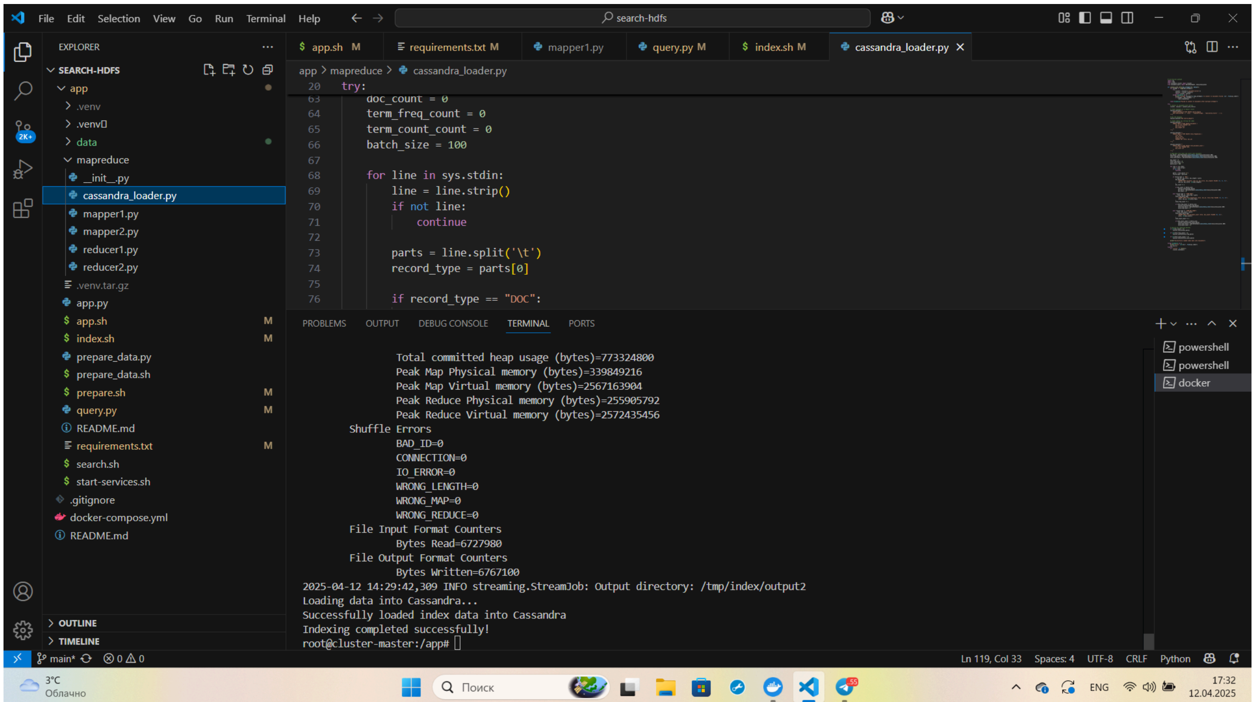

File Edit Selection View Go Run Terminal Help
search-hdfs
EXPLORER
SEARCH-HDFS
app
data
2855741_AMD_mobile_platform.txt
2869726_XCircuit.txt
2924800_XBase++.txt
2931262_Xubuntu.txt
2969420_Xenohyla.txt
3001778_A_Black_&_White_Night_Live.txt
3008420_Xintai.txt
3010402_Abdool_Razack_Mohamed...
3025051_Xing.txt
3030811_Aag_(2007_film).txt
3208750_X10_Wireless_Technology.txt
3215637_Xavier_Ortiz.txt
3230864_Xavier_Escudé.txt
3237381_ASAP.txt
3363082_XHCJE-TDT.txt
3367385_Xavier_Dirceu.txt
3375522_Xavix.txt
3390416_Xiao_(flute).txt
3441475_X-15_(film).txt
3448443_Xanadu_(Rush_song).txt
3499174_Xiled_to_Infinity_and_One.txt
3517745_X_Division.txt
3531954_Xunhua_Salar_Autonomou...
3545257_XED.txt
3548475_Xiomara_Laugart.txt
3549644_APOEL_Nicosia.txt
3553124_X-ray_reflectivity.txt
3573117_X-ray_crystal_truncation_ro...
OUTLINE
TIMELINE
app.sh M
requirements.txt M
mapper1.py
query.py M
3517745_X_Division.txt U
index.sh M
app > data > 3517745_X_Division.txt
becoming one of the best wrestlers in the division. Both Daniels and Styles disliked Joe, despite having had a feud of their own. ==Tournaments== TNA maintains three different styles of tournament referred to as "X Cup" tournaments. The Super X Cup tournament is a standard single-elimination tournament featuring one-on-one matches. The Americas X Cup tournament was a team-format points-based tournament featuring two teams of four wrestlers each, with each team representing a respective country that most or all of the wrestlers are from. Members of the team competed in a variety of matches, including singles matches and tag team matches, which accrued points for their side. The World X Cup tournament was an expansion on the Americas X Cup, in which four teams of four wrestlers competed. In the World X Cup, TNA always hosts a home team for the United States, with other countries such as Japan, Mexico, and Canada being represented by either TNA-contracted wrestlers or wrestlers from a promotion that TNA has a partnership agreement with. X Division wrestlers are generally the only TNA wrestlers that compete in the TNA X Cup Tournaments. The first such tournament was the TNA 2003 Super X Cup Tournament, which was won by Chris Sabin. ===TNA X Division Championship Tournament (2009)=== The tournament was the result of a match for the TNA X Division Championship at Final Resolution between Eric Young and Sheik Abdul Bashir ending in a controversial fashion, with Young winning the championship thanks to the referee's help. Management Director Jim Cornette stripped Young of the belt and announced the tournament to crown the new champion. The tournament final took place at Genesis. ===TNA X Division Championship #1 Contender Tournament (2011)=== On the January 27, 2011, edition of Impact!, TNA started a tournament to determine a new number one contender for the TNA X Division
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
root@cluster-master:/app# ./search.sh "Super X Cup Tournament"
This script will include commands to search for documents given the query using Spark RDD
Searching for: Super X Cup Tournament
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-0c7eb1a8-d770-4abe-abb6-2e96ad178a32;1.0
confs: [default]
found com.datastax.spark#spark-cassandra-connector_2.12;3.1.0 in central
found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.1.0 in central
found com.datastax.oss#java-driver-core-shaded;4.12.0 in central
found com.datastax.oss#native-protocol;1.5.0 in central
found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
found com.typesafe#config;1.4.1 in central
found org.slf4j#slf4j-api;1.7.26 in central
found io.dropwizard.metrics#metrics-core;4.1.18 in central
found org.hdrhistogram#HdrHistogram;2.1.12 in central
found org.reactivestreams#reactive-streams;1.0.3 in central
found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
powershell
powershell
docker
main* 0 0
Ln 1, Col 4404 (22 selected) Spaces: 4 UTF-8 CRLF Plain Text
3°C
В осн. солнечно
Поиск
ENG
17:16
12.04.2025


PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
==
TOP 10 SEARCH RESULTS
==
1. X_Division (ID: 3517745, Score: 17.8962)
2. Xande_Silva (ID: 43589266, Score: 13.6063)
3. Ximena_Ros (ID: 63158494, Score: 13.4498)
4. Xander_Schauffele (ID: 51911552, Score: 13.2957)
5. Xu_Jiamin (ID: 59128773, Score: 11.5962)
6. APOEL_Nicosia (ID: 3549644, Score: 10.5198)
7. Xavi_Llorens (ID: 55607189, Score: 10.0040)
8. Xiao_Guodong (ID: 15806433, Score: 9.9760)
9. Xu_Yifan (ID: 27861895, Score: 9.7918)
10. Xagra_United_F.C. (ID: 17968524, Score: 9.5978)
==
25/04/12 14:17:33 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/12 14:17:33 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/12 14:17:33 INFO BlockManagerInfo: Removed broadcast_5_piece0 on cluster-master:36637 in memory (size: 6.5 KiB, free: 366.3 MiB)
25/04/12 14:17:33 INFO BlockManagerInfo: Removed broadcast_5_piece0 on cluster-slave-1:36329 in memory (size: 6.5 KiB, free: 366.3 MiB)
25/04/12 14:17:33 INFO BlockManagerInfo: Removed broadcast_5_piece0 on cluster-slave-1:40791 in memory (size: 6.5 KiB, free: 366.3 MiB)
25/04/12 14:17:33 INFO YarnClientSchedulerBackend: Interrupting monitor thread
17:17
12.04.2025



File Edit Selection View Go Run Terminal Help
search-hdfs
EXPLORER
SEARCH-HDFS
app
.venv
.venv
data
mapreduce
__init__.py
cassandra_loader.py
mapper1.py
mapper2.py
reducer1.py
reducer2.py
.venv.tar.gz
app.py
app.sh
index.sh
prepare_data.py
prepare_data.sh
prepare.sh
query.py
README.md
requirements.txt
search.sh
start-services.sh
.gitignore
docker-compose.yml
README.md
OUTLINE
TIMELINE
app.sh M
requirements.txt M
mapper1.py
query.py M
index.sh M
cassandra_loader.py
app > app.sh
venv-pack -o .venv.tar.gz
hdfs dfs -mkdir -p /index/data
hdfs dfs -mkdir -p /user/root
hdfs dfs -chmod -R 777 /user/root
Run the indexer
echo "Running indexer..."
bash index.sh /index/data
Run the ranker
echo "Running sample search..."
bash search.sh "this is a query!"
EOF
chmod +x /app/app.sh
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
root@cluster-master:/app# ./index.sh /index/data
This script include commands to run mapreduce jobs using hadoop streaming to index documents
Input file is:
/index/data
Copying data to HDFS...
Deleted /index/data/1000197_XHVE-FM.txt
Deleted /index/data/10211542_A._Maceo_Walker.txt
Deleted /index/data/10216019_Xela_Arias.txt
Deleted /index/data/10437846_A_Huguenot_on_St._Bartholomews_Day.txt
Deleted /index/data/10627402_X3_train.txt
Deleted /index/data/10917985_Xingliao.txt
Deleted /index/data/10927998_Xaltocan_Tlaxcala.txt
Deleted /index/data/10983606_Xiaoshangqiao.txt
Deleted /index/data/1104168_X47.txt
Deleted /index/data/110918_Xenia_Illinois.txt
Deleted /index/data/11116129_XHNOE-FM.txt
Deleted /index/data/11116350_XHAHU-FM.txt
Deleted /index/data/11116435_XEFE-AM.txt
Deleted /index/data/11116789_XHWL-FM.txt
Deleted /index/data/11128761_Xylaria_mali.txt
Deleted /index/data/11174498_A_Night_Under_the_Dam.txt
powershell
powershell
docker
main*
Enin Dmitriy (4 days ago) Ln 29, Col 15 (11 selected) Spaces: 4 UTF-8 LF Shell Script
3°C
Облачно
Поиск
ENG
17:09
12.04.2025




File Edit Selection View Go Run Terminal Help
search-hdfs
EXPLORER
SEARCH-HDFS
app
.venv
.venv0
data
mapreduce
__init__.py
cassandra_loader.py
mapper1.py
mapper2.py
reducer1.py
reducer2.py
.venv.tar.gz
app.py
app.sh
index.sh
prepare_data.py
prepare_data.sh
prepare.sh
query.py
README.md
requirements.txt
search.sh
start-services.sh
.gitignore
docker-compose.yml
README.md
OUTLINE
TIMELINE
app.sh M
requirements.txt M
mapper1.py
query.py M
index.sh M
cassandra_loader.py
app > mapreduce > cassandra_loader.py
try:
doc_count = 0
term_freq_count = 0
term_count_count = 0
batch_size = 100
for line in sys.stdin:
line = line.strip()
if not line:
continue
parts = line.split('\t')
record_type = parts[0]
if record_type == "DOC":
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Total committed heap usage (bytes)=773324800
Peak Map Physical memory (bytes)=339849216
Peak Map Virtual memory (bytes)=2567163904
Peak Reduce Physical memory (bytes)=255905792
Peak Reduce Virtual memory (bytes)=2572435456
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=6727980
File Output Format Counters
Bytes Written=6767100
2025-04-12 14:29:42,309 INFO streaming.StreamJob: Output directory: /tmp/index/output2
Loading data into Cassandra...
Successfully loaded index data into Cassandra
Indexing completed successfully!
root@cluster-master:/app#
powershell
powershell
docker
main* 0 0
Ln 119, Col 33 Spaces: 4 UTF-8 CRLF Python
3°C
Облачно
Поиск
ENG
17:32
12.04.2025