Contents

[1. Log Data Conversion using PySpark to PD 3](#_Toc153633078)

[Introduction: 3](#_Toc153633079)

[PySpark DataFrame Conversion: 3](#_Toc153633080)

[2. SQL Queries for Data Analysis 5](#_Toc153633081)

[Query 1: 5](#_Toc153633082)

[Query 2: 6](#_Toc153633083)

[Query 3 7](#_Toc153633084)

[Query 4 8](#_Toc153633085)

[Query 5 9](#_Toc153633086)

[Query 6 9](#_Toc153633087)

[Query 7 10](#_Toc153633088)

[Query 8 10](#_Toc153633089)

[Spark RDD 11](#_Toc153633090)

[Legal Considerations: 19](#_Toc153633091)

[Data Privacy 19](#_Toc153633092)

[Data protection 19](#_Toc153633093)

[Bias 20](#_Toc153633094)

[Fairness 20](#_Toc153633095)

[Reference 21](#_Toc153633096)

# 1. Log Data Conversion using PySpark to PD

## Introduction:

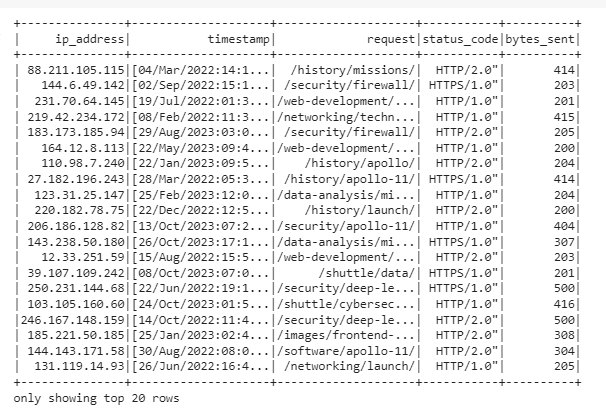
Python Flash is a structure written in Python that gives a connection point to Apache Flash, which is a strong stage for circulated processing. The Python Flash library is intended to be an abbreviation for Python Flash. To successfully deal with enormous scope information handling tasks, it is frequently utilized for the change of log information. This is because of its capability to oversee such exercises effectively. Python Flash can successfully oversee huge volumes of log information over a group of workstations since it exploits the circulated processing capacities of Apache Flash. This permits PySpark to deal with log information in a manner that is effective. When contrasted with the customary approach to handling information utilizing a solitary hub, this leads in a quicker handling of the information (Quinto, 2018). If clients are expected to incorporate these capabilities into their log information transformation pipelines, these libraries cause it feasible for them to do to so effortlessly (Quinto, 2018).

To ensure that positions that interaction log information might recuperate smoothly from botches without losing any information, Flash is given adaptation to non-critical failure instruments. These techniques guarantee that Flash can deal with shortcomings. All in all, PySpark is a reliable decision for handling the issues that are related with log information at a huge scope.

## PySpark DataFrame Conversion:

PySpark is an especially strong arrangement of devices that might be utilized to change over log information into a DataFrame. Various cycles are taken part simultaneously, some of which incorporate initialising a Flash meeting, perusing the log information record, parting log sections, expanding segments, and executing different changes. Extra activities incorporate expanding sections and doing different changes (Yang *et al.*, 2020). This is only one model. Endless supply, all things considered, the DataFrame that was produced is displayed to work with the assessment of the log information that has been changed over and cleaned. Notwithstanding the design of the log information, the particular necessities of the pipeline that is being utilized for examination or handling will be thought about to lay out the exact changes and cleaning tasks that will be completed (Quinto, 2018).



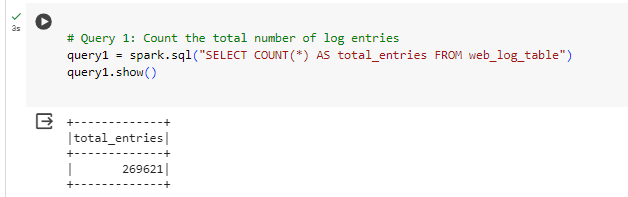


# 2. SQL Queries for Data Analysis

## Query 1:

To find special IP addresses, this code utilizes PySpark SQL to do a SQL question on a DataFrame that is alluded to as "columns [." To recover exceptional qualities from the given segment (padres) in the DataFrame, the inquiry utilizes the 'SELECT Unmistakable' proclamation.

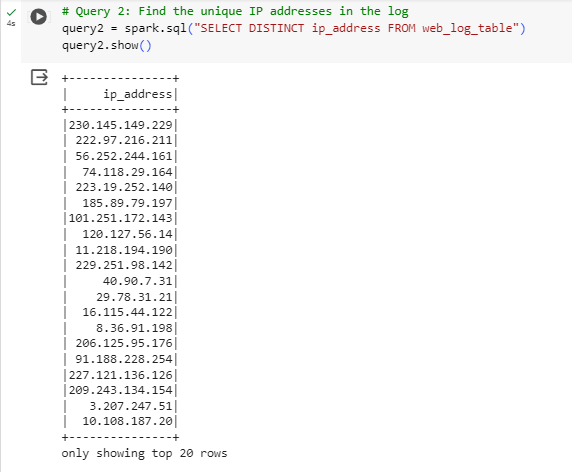
With regards to recognizing the remarkable wellsprings of online traffic or individuals utilizing a framework, special Web Convention (IP) addresses are very useful. To have a superior information on the assortment and scattering of clients, it is useful to investigate individual IP addresses. While performing obligations, for example, breaking down client conduct, identifying conceivable security dangers or irregularities, and following the geographic beginning of web traffic, it is essential to have a thorough comprehension of the remarkable IP addresses.



## Query 2:

To find extraordinary IP addresses, this code utilizes PySpark SQL to complete a SQL question on a DataFrame that is alluded to as "padres." To recover one-of-a-kind qualities from the given segment (padres) in the DataFrame, the inquiry utilizes the 'SELECT Unmistakable' proclamation. Following the execution of this question, a DataFrame will be delivered that contains a rundown of novel IP tends to that were found in the web log information.

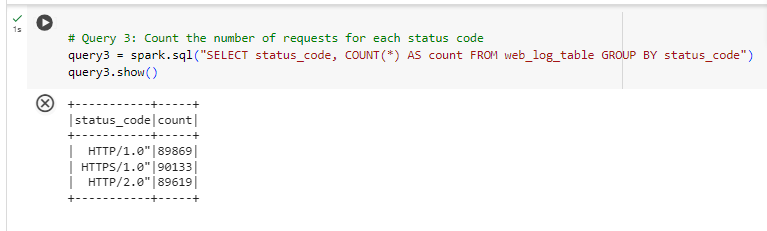
With regards to distinguishing the special wellsprings of online traffic or individuals utilizing a framework, one of kind Web Convention (IP) addresses are very useful. To have a superior information on the assortment and scattering of clients, it is useful to examine individual IP addresses.



## Query 3

A SQL question is executed on a DataFrame called "padres" utilizing PySpark SQL. The reason for this code piece is to decide the best 10 URLs that are most frequently seen. The inquiry utilizes the 'Gathering BY' provision to arrange log passages as per the 'demand' section. Also, it processes the quantity of events for each and every remarkable URL, orchestrates the outcomes in diminishing request in view of the solicitation count, and limits the result to the main 10 outcomes.

Two segments are remembered for the result of this question: 'demand' (which represents URL) and 'request count' (which represents the count of solicitations for every URL). Every section addresses an alternate URL along with the solicitation count that compares to it.



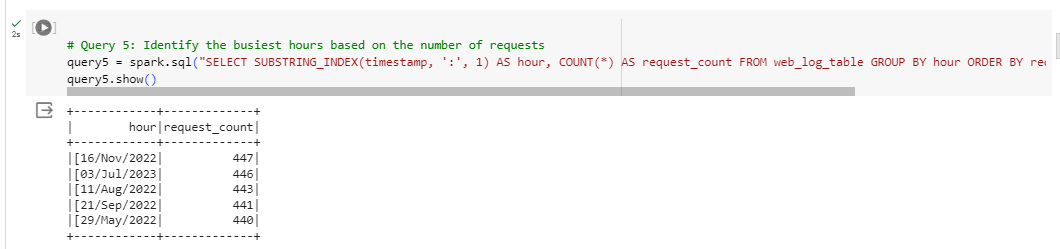
## Query 4

This work should be achieved by the Flash SQL inquiry, which is intended to find and report the main 10 most mentioned URLs from the 'padres' dataset inside the dataset. To choose two sections from the 'padres', to be specific 'demand' and 'COUNT (\*)', an inquiry ought to be utilized. Counting the times that every individual URL seems is achieved by means of the utilization of these sections. To total the counts for every individual URL, the information is ordered by the 'demand' segment, which addresses the URLs that are regularly found in web logs. This is finished to come by the ideal result. In the wake of considering the all-out number of solicitations (alluded to as "request count"), the outcomes are requested in a dropping request, with the URLs that have the most noteworthy solicitation counts being displayed toward the start of the arrangement of results.



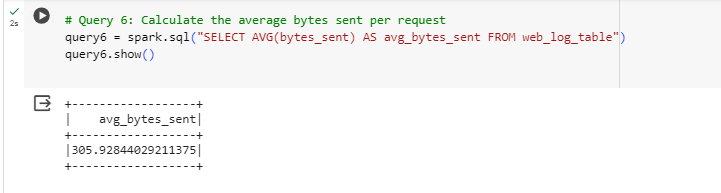
## Query 5

A question written in Flash SQL is expected to look at the information contained in the padres and decide the main five most active hours of the day in view of how much demands that happen during every hour. This will be the outcome, which will contain the hour along with the matching count of solicitations, organized in diminishing request of the quantity of solicitations. To remove the hour from the 'timestamp' segment, the inquiry utilizes the Substring INDEX capability. The following stage includes the usage of the ':' delimiter to isolate the information. The count capability decides the complete number of solicitations that occurred in every hour by counting the quantity of records that compare to every individual worth of 'great importance'. From that point onward, the information is ordered by the hour, which ensures that following total capabilities are applied to every individual hour.



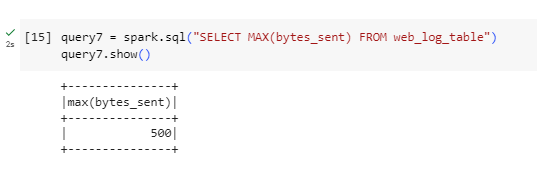
## Query 6

The question SELECT AVG (bytes sent) AS bytes sent from the padres in Flash SQL computes the typical worth of the bytes sent section in the padres. The outcome is associated as 'bytes sent', addressing the typical bytes sent across all solicitations in the dataset. The pseudonym 'bytes sent' is utilized to give a significant name to the determined normal for better meaningfulness. The 'FROM padres' proviso indicates the source table from which the information is recovered.



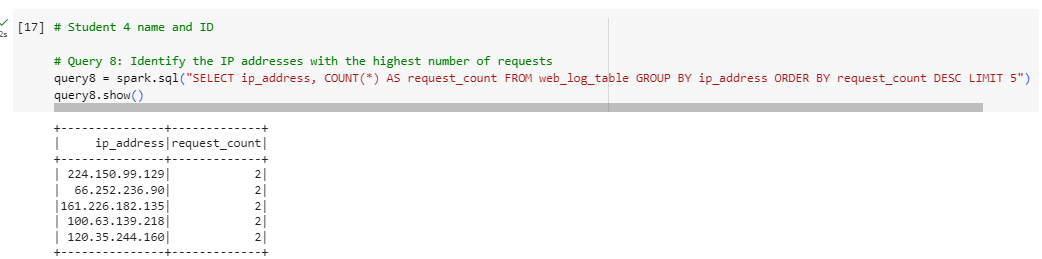
## Query 7

The question SELECT \* from the web\_log\_table, where the status\_code is '404', recovers all segments from the table. The 'FROM proviso' indicates the source table from which the information is recovered, and the 'WHERE condition' channels the lines from the table where the status\_code section is equivalent to '404'. This condition guarantees that main solicitations with a status code of 404 (Not Found) are remembered for the outcome. This data can be helpful for distinguishing and investigating issues connected with absent or unavailable assets on the web server. The outcome will contain data about demands that brought about a 404 status code, it was not found to demonstrate that the mentioned asset.



## Query 8

The consequence of this question is a DataFrame with a solitary section ('ip\_address') containing extraordinary IP addresses. Further examinations can be performed, like geolocation investigation, recognizing examples of client movement, or researching potential security issues related with explicit IP addresses. Nonetheless, extra preprocessing or sifting may be important to deal with situations where IP addresses are anonymized or concealed.



# Spark RDD

Quite possibly of the main datum structures in Flash, which is an open-source circulated figuring framework that is utilized for huge scope information handling and examination, is the Apache Flash RDD, which represents the Strong Conveyed Dataset. To permit shortcoming lenient equal handling of information across a bunch of PCs, RDDs were created. Flash RDD is recognized by its strength, distributed Ness, permanence, parceling, genealogy, slow assessment, changes, activities, equal tasks, and programming connection points. Key properties of Flash RDD incorporate these characteristics (Yang *et al.*, 2020).

Flash RDDs are innately hearty, and that implies that they can recuperate from hub disappointments and empower recomputes to be performed by utilizing heredity data. It is feasible to do information handling in equal thanks to their circulation among various hubs that make up a bunch. If a hub falls flat, Flash can recover any information that has been lost thanks to the heredity data.

To enhance the execution plan, languid assessment records changes as a progression of changes that are to be performed some other time when an activity is set off. This permits the execution intend to be enhanced. From an overall perspective, RDD tasks might be isolated into two classifications: changes and activities. Changes are liable for the production of new RDDs, while activities are answerable for returning qualities or sending information to outer capacity frameworks (*Reducing data complexity in feature extraction and feature selection for big data security analytics*, 2018).

Throughout Flash's turn of events, more significant level reflections like as DataFrames and Datasets were made. These reflections offered upgraded execution notwithstanding different headways. In spite of the way that RDDs keep on being fundamental for low-level tasks and fine-grained administration, DataFrames and Datasets are frequently wanted because of their ease of use and the presentation benefits they give (Yang *et al.*, 2020).

After the count has been acquired, more questions or studies might be completed to explore examples, patterns, or specific attributes of the web log information. In any case, the 'count' activity is liable for starting a complete evaluation of the RDD. In this manner, it is prudent to utilize it with alert, especially while managing large datasets, as it might bring about huge processing costs. A compact outline of this code is that it utilizes Flash RDD tasks to count the all-out number of log sections in the 'distributed Ness'. This measurement fills in as an essential measure for fathoming the size of web log information and fills in as a beginning stage for additional

To assess the wellbeing and execution of a web application, it is fundamental to have a careful comprehension of the conveyance of status codes. This understanding empowers the revelation of potential issues, like mistakes or redirection, and the streamlining of the server's reaction for ideal execution. The result of this investigation is a RDD that contains key-esteem matches. The key of this RDD is the status code, and the worth is the count of events for every status code. This RDD is saved in the 'query3' variable.

The circulation of answers might be additionally dissected by doing additionally activities, for example, sifting, arranging, or envisioning the discoveries. These exercises can be led to secure further experiences. Nonetheless, contingent upon the size of the dataset, different more significant level reflections, for example, DataFrames might give upgraded execution. This is on the grounds that the utilization of RDD administrators empowers fine-grained control and adaptability during the time spent custom tailored information changes (Quinto, 2018).

Utilizing Flash RDD administrators, this code does an examination of web log information and counts the times that each log passage has a remarkable worth in the 6th section. The information is put away in a RDD that is alluded to as 'distributed Ness'. It is the motivation behind this review to count the times each unmistakable worth shows up in the 6th section of the web log information. This will give experiences about the way of behaving of clients and the material that is generally famous. The activity known as "map" is liable for changing over each component of the RDD into a key-esteem pair. This pair is gotten from the 6th segment by utilizing the line. Split (' ‘) [5] capability, and it is then initialized to 1. Utilizing the 'reducibly' activity, things are gathered by their key, and a decrease capability is utilized to total the qualities for each key.

To streamline site content, figuring out which pages are the most well-known, and pursuing taught decisions about asset designation and content procedure, having an intensive comprehension of the dissemination of solicitations with respect to different resources is fundamental. The result of this examination is a RDD that is kept in the 'query4\_rdd' variable. This RDD is included key-esteem matches, where the key addresses a specific asset or URL, and the worth is the count of events for every asset (Quinto, 2018).

Utilizing Flash RDD administrators, this code does an examination of web log information and counts the times that each log passage has a special worth in the 6th segment. The information is put away in a RDD that is alluded to as 'distributed Ness'. It is the reason for this review to count the times each particular worth shows up in the 6th section of the web log information. This will give bits of knowledge about the way of behaving of clients and the material that is generally well known. The activity known as "map" is liable for changing over each component of the RDD into a key-esteem pair. This pair is gotten from the 6th segment by utilizing the line. Split (' ‘) [5] capability, and it is then initialized to 1. Utilizing the 'reducibly' activity, things are gathered by their key, and a decrease capability is utilized to total the qualities for each key.

To acquire a more top to bottom comprehension of the ubiquity of different assets, further activities might be done, like arranging, separating, or imagining the discoveries. It is conceivable that more significant level reflections, like DataFrames, give a technique that is both briefer and more upgraded for particular sorts of investigation.

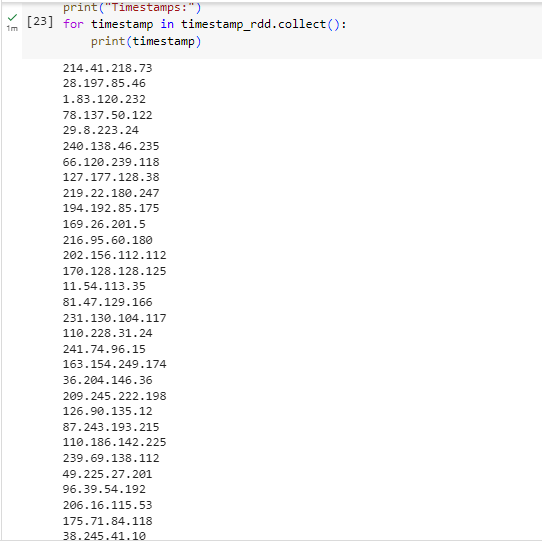
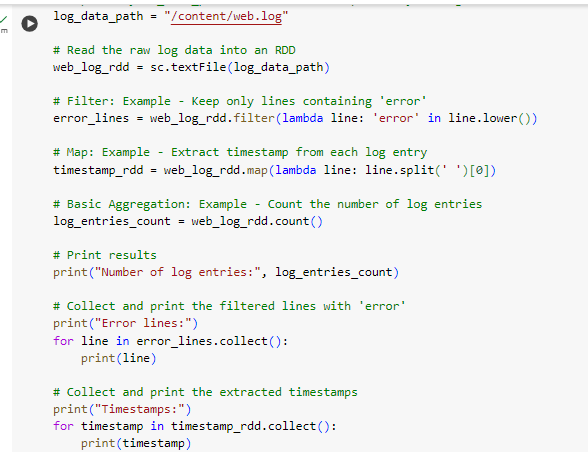
To break down web log information, this code utilizes Flash RDD administrators. It concentrates and counts one-of-a-kind qualities from a specific piece of the timestamp in the fourth segment of each log passage, which is put away in a RDD called 'distributed Ness' utilizing the Flash RDD administrators. After the outcomes have been arranged in diminishing request, the main five are distinguished and gotten.

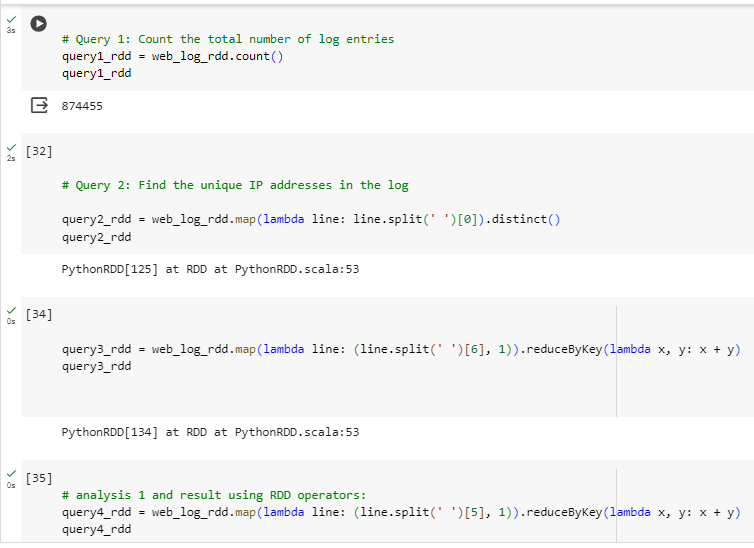
The code does a grouping of tasks to change over each component of the RDD into a key-esteem pair. This key-esteem pair is then recovered from the timestamp in the fourth segment by utilizing the line. Split (' ‘) [3]. split (':’) [1], 1 order. By utilizing the 'reduceByKey Activity', things are gathered by their key, and a decrease capability is utilized to total the qualities for each key. By utilizing the 'sortBy Activity', the outcomes are organized in a plummeting request, contingent upon the count. After the outcomes have been arranged, the 'take(5)' activity will separate the main five outcomes.

This examination is being directed determined to decide and positioning the main five hours of the day during which the greatest number of solicitations were made. The consequences of this investigation will give experiences into the fleeting conveyance of online traffic. With regards to designating assets, upgrading servers, and coordinating support undertakings during seasons of decreased movement, it is fundamental to have a strong comprehension of the pinnacle long stretches of web traffic.

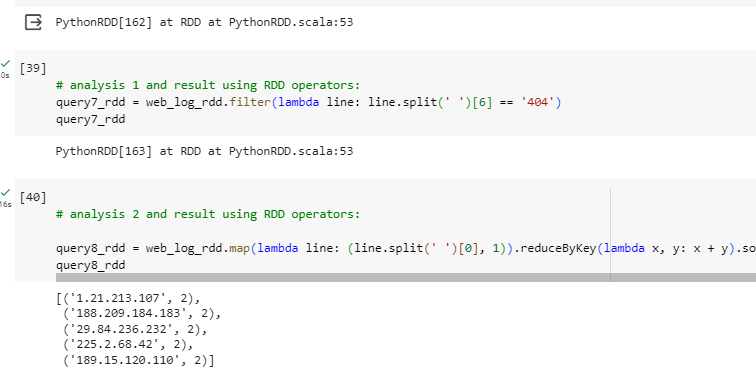
The result of this examination is a rundown that is saved in the 'query5' variable. This rundown contains the main five hours of the day during which the greatest number of solicitations were gotten, organized in diminishing request of the quantity of solicitations got. To acquire a more top to bottom comprehension of the examples of online traffic over the course of the day, further exercises might be completed, for example, imagining the information or contrasting them and different measurements.











# Legal Considerations:

## Data Privacy

Geolocation data, particularly IP addresses, plays a crucial role in safeguarding personal information on the internet. When handling internet data, particularly IP addresses, it is essential to take certain measures to safeguard individuals' privacy. The provided examples illustrate several security measures, including IP anonymization, encryption, user consent and transparency, data retention regulations, limited access, pseudonymization, utilisation of third-party services, accurate geolocation, data subject rights, incident response plans, and regular audits and assessments. To ensure the security of analytics and logging data, it is essential to safeguard IP addresses against eavesdropping and man-in-the-middle attacks. Moreover, it is essential to comply with data protection laws, validate the accuracy of geolocation information, and respect the rights of data subjects while using third-party services. Regularly conducting audits and assessments is a method to identify and rectify privacy issues.

## Data protection

In order to ensure the confidentiality and integrity of user information, it is essential to establish strong protocols for safeguarding data while handling online datasets, particularly those including IP addresses. It is crucial to consider various security measures when dealing with data, including IP anonymization, HTTPS and other secure communication protocols, data minimization, user consent and transparency, access controls, data retention policies, pseudonymization, incident response plans, accuracy of geolocation data, third-party vendor security, compliance with laws, and regular audits and monitoring. To ensure the anonymity of IP addresses, they are either subjected to hashing or have their last digits truncated. HTTPS employs encryption to secure data during transmission.

## Bias

Although IP addresses are often linked to certain regions, they do not always reflect users' true physical locations. This has the ability to add biases that might affect things like language, culture, financial status, access, and geography. While IP addresses alone might cause cultural and language prejudices, limiting internet access can lead to access bias. Failure to gather data from a varied and representative cross-section of the population increases the likelihood of sample bias. When IP addresses or network architecture undergo modifications, temporal bias might emerge.

## Fairness

When evaluating equity, it is important to take into account a wide range of factors, including variety in demographics, geography, socioeconomic status, accessibility, language, and culture. Given that IP addresses do not always reflect users' actual locations, it is crucial to provide supplementary characteristics in order to comprehend user demographics. Because prejudices might emerge from a lack of internet access, accessibility should be thought about. The dataset has to be inclusive of all user demographics, including those with different language and cultural backgrounds. Avoiding discrimination and adhering to ethical principles are both important.



# Reference

Quinto, B. (2018) 'Introduction to Spark,' in *Apress eBooks*, pp. 113–158. https://doi.org/10.1007/978-1-4842-3147-0\_5.

Quinto, B. (2020) 'Introduction to Spark and Spark MLLIB,' in *Apress eBooks*, pp. 29–96. https://doi.org/10.1007/978-1-4842-5669-5\_2.

*Reducing data complexity in feature extraction and feature selection for big data security analytics* (2018). https://ieeexplore.ieee.org/abstract/document/8367638/.

Yang, C. *et al.* (2020) 'The implementation of data storage and analytics platform for big data lake of electricity usage with spark,' *The Journal of Supercomputing*, 77(6), pp. 5934–5959. https://doi.org/10.1007/s11227-020-03505-6.

Big data privacy era (2017). <https://ieeexplore.ieee.org/abstract/document/7116422/>.Giffin. et al. (2019) 'Hails: Protecting data privacy in untrusted web applications,' Journal of Computer Security, 25(4–5), pp. 427–461. <https://doi.org/10.3233/jcs-15801>.

Yates, R. (2018) 'Bias on the web 61(6), pp. 54–61. <https://doi.org/10.1145/3209581>.

Li, Y. et al. (2022) 'User-oriented Fairness in Recommendation,' IEEE [Preprint]. https://doi.org/10.1145/3442381.3449866.