What are the policies and processes that should be put in place to ensure the **privacy and security** of data in data engineering projects?

1. Chong Jing Yung
2. Ong Weng Kai
3. Ryan Kho Yuen Thian
4. Sim Hong Li
5. Thong Cheng How
6. Yong Zee Lin

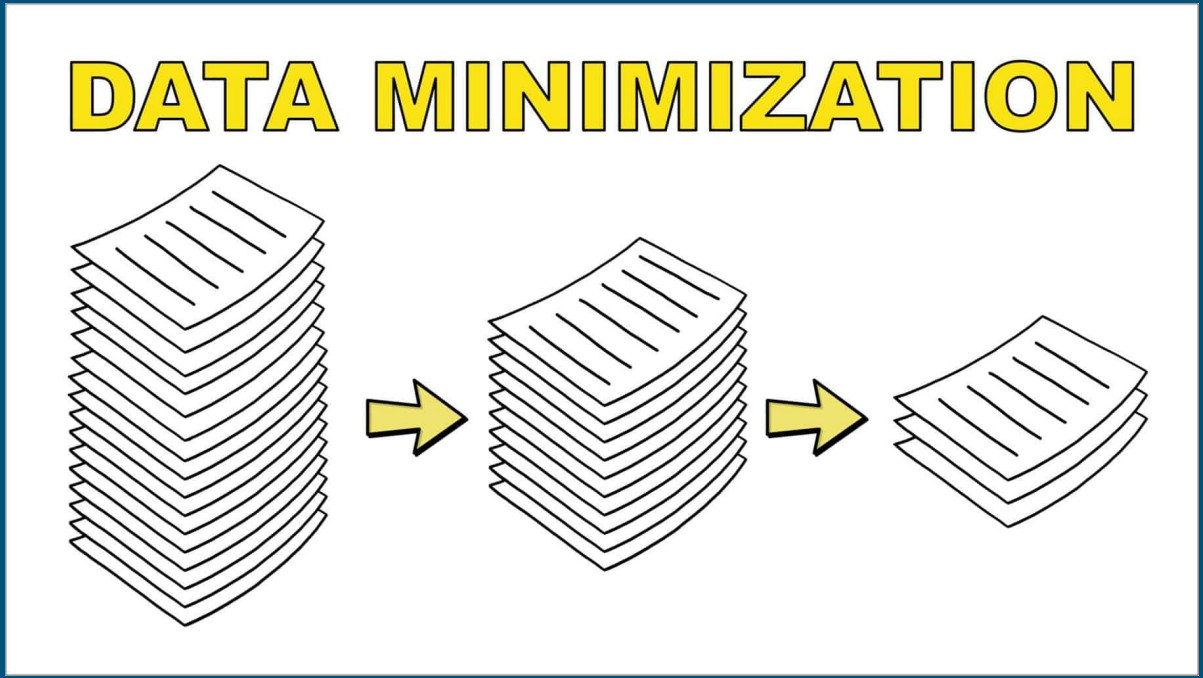
G23

Aspect of Privacy & Security

1. Ryan Kho Yuen Thian
2. Ong Weng Kai
3. Sim Hong Li

Data Minimization when Web Scraping Data

By: Ryan Kho Yuen Thian



What is Data Minimization

- 1 of the Important Data Protection Principles
- Demands that you gather and keep only the bare minimum of data necessary for delivering a product or service
- Pioneered by the European Union's General Data Protection Regulation (GDPR)
- Do not collect personal data unless it directly benefits you.
- One must also decrease the amount of data they already have
- Simplifies Personal Data Protection Act (PDPA) adherence by minimizing the volume of data that organisations need to manage, safeguard and ensure protection for

(WinZip, 2023; Oh, 2023)

What European Data Protection Supervisor says about Data Minimization

Data minimization

The principle of “data minimisation” means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. They should also retain the data only for as long as is necessary to fulfil that purpose. In other words, data controllers should collect only the personal data they really need, and should keep it only for as long as they need it.

The data minimisation principle is expressed in Article 5(1)(c) of the GDPR and Article 4(1)(c) of Regulation (EU) 2018/1725, which provide that personal data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".

(EDPS, 2024)

Related Works

Sentiment Analysis on Youtube Comments: A brief study (Akhtar, 2019)

All said about the kind of data, now the **ID of the video** is **Y-XHMlaJL-s.**
With the help of this ID of video, that YouTube uniquely assigns, we will download the comments and which will be saved in a **json** format. Also, while downloading the comments, we will be asked to state the number of comments we would like to download. In my case, I have kept data small i.e, **Dataset Size 50.** The output saved file name is **'football_video.json'.**
The file have extracted 4 key value pair, namely, [cid, text, author, time]. Here, cid is the Comment ID.

jupyter football_video.json 30 minutes ago

File Edit View Language


```
1 {"cid": "UgzmE_SaYibGc6aRgul4AaABAg", "text": "Btsport should use this when the final comes on live f
good this is. Would make the pre-game show even better if they could use this", "time": "58 minutes a
2 {"cid": "UgzQVLtsKzaBnzIEpCx4AaABAg", "text": "Can anyone tell me the first Track of the Video?😃", '
"author": "Frank Kraemer"}
3 {"cid": "Ugzd73H5MfqY6E_9kWZ4AaABAg", "text": "COYR", "time": "8 hours ago", "author": "BIG MAC HD"}
4 {"cid": "Ugze-DBO5m3h6A3Plg54AaABAg", "text": "Music please", "time": "11 hours ago", "author": "HFS
5 {"cid": "Ugzlx4jttfOuvPoJ54t4AaABAg", "text": "I am a liverpool fan i believe well  beat tottenham",
"Micheál O Kelly"}
```

Related Works (Con't)

YouTube Video Analysis (Bachubhay et al., 2021)

Figure 3: CSV Output for an Example Video's Comments

cid	author	text	time	votes
UgyhLWFHthiP7	juxxin17u11	Biffa which DLC	2021-03-26T12:	1
Ugxt1YgV3FCpc	Finshark007_rob	Biffs casually wa	2021-03-14T11:	0
Ugw-cTxbAfKUx	Andrew Proctor	<a href="https://\	2021-03-02T05:	0
UgyArXjnKpKXY	Jay Murray	Already mention	2021-01-28T15:	2
UgyMBbdjZ4DK	LikeableKiwi	if you dont want	2021-01-02T09:	1
UgzNmOt5onhQ	Ivan Semanco	like ;-] greit viteo	2020-12-28T19:	0
UgwzP0s1kx8Tx	Daniel Coffey	I am enjoying a a	2020-12-27T07:	0
Ugyv4rNb77Dq-	Jim Frost	Trams are limite	2020-12-20T19:	0
UgyfkBrCYzvucf	Stoyan Stoitsev	Should get some	2020-12-12T10:	0

Sentimental Analysis on YouTube Scrapped Data

(G & P, 2022)

	Comment	Likes	Time	UserLink	user
0	[LIMITED STOCK] Buy a cod	175	1 month a	https://www.youtube.com/channel/UC	
1	Dil se shukriya Harry sir ..lo	8.6K	1 year ago	https://w	WRESTLING STAR-WWE
2	7:06:46I have created my fi	15	1 day ago	https://w	Jajati Satpathy
3	6:52:00	3	2 days ago	https://w	Sabnam Laskar
4	1:37:39	1	19 hours a	https://w	Arihant Jain
5	for printing the "object is c	4	2 days ago	https://w	Mitali Sawarkar
6	n = int(input("enter a num	4	2 days ago	https://w	Basic of computer science
7	The guy is working hard ev	1.2K	1 year ago	https://w	Rocker Techs
8	Whenever I thought about	29	8 days ago	https://w	Sahil Jadhav
9	I was trying to learn pytho	28	2 weeks a	https://w	Shashwat Kumar
10	5:13:34	3	1 day ago	https://w	HARSH DESHMUKH
11	Sir I'm in class 8th right no	4	1 day ago	https://w	coder003
12	5:05:20a = input("Enter use	2	2 days ago	https://w	Noman Khan
13	The only who is providing	226	1 year ago	https://w	Krishnan Kundan
14	I am from Bangladesh and	2	1 day ago	https://w	faiyaz mahmud
15	1:37:06 (day 3)	6	2 weeks a	https://w	Vanshit
16	5:51:41	3	3 days ago	https://w	SHASHANK SINGH
17	9:14:56a function is used f	3	3 days ago	https://w	Abhijeet Redekar
18	Guys' let's not forget he is	893	1 year ago	https://w	Sayyed Abid

Figure 5:Comments received

How I applied Data Minimization

comment_text	comment_author	votes	dislikes	replies	id
I like how codm p...	@--FlameZ--	51	0	3	UgyTUAWhSyj0QI6Pz...
Please add the we...	@Angel--AR-15	7	0	NULL	Ugy_rJN270SWTrHvt...
Blitz Is Amazing ...	@gaminguceyt	2	0	NULL	Ugx82pUfNqR-GWaig...
I'm just saying i...	@jaevaldez9650	1	0	NULL	UgxtOCQDBmdL2bZ7p...
Button to block m...	@Martin-zb4rk	3	0	3	UgzJsAoP-3kzZmLh9...
Valentine's Day i...	@morganmeliora2603	1	0	NULL	UgxABlusdyDN8i5Lh...
Hello to the Acti...	@M-ev1zz	0	0	NULL	Ugzlwo2nHEgOUc2yB...
Would be nice to ...	@KingPepper41	4	0	NULL	UgxSjyCV-jwbpv7gX...
Still waiting for...	@huntervincent579	0	0	NULL	Ugwwj8DJBfWisc3Rh...
the map is beauti...	@SlayinSiren	0	0	NULL	Ugw1tzBaqOJ-snHIc...
WHEN WILL YOU RET...	@K.A.R.A.U	0	0	NULL	UgwEkadtPNo1UeM6o...
Where siren!!	@hanzxianvlog	0	0	NULL	UgzSZXHRUJrbRt4Io...
We need an old Ba...	@alirezarzaqzade4484	2	0	1	UgxCia6m32Y1ovipM...
Give the opportun...	@zeroalfa016	0	0	NULL	Ugx5i9ryxf4hFst1W...
Its been a while ...	@hardy352	0	0	NULL	UgzQltfcw3tPmJShP...
Mythic fennec pls	@iPadKid1324	0	0	NULL	UgwodyJpCH8EHLw8U...
Bring back MEMNOS	@SINfromPL	1	0	NULL	UgwBGqSE0db95o5AT...
Pls add br practi...	@AidenFunnyShorts...	0	0	NULL	UgzOhbhSqRp0uZdFi...
Battle Pass Seaso...	@COD_PRO8L3M	1	0	NULL	UgyZLOEdRt9nA6PDX...
Thanks for the fr...	@abolfazl648	0	0	NULL	UgyKmSKzNhBu-YR7z...

only showing top 20 rows

Dataset formed from Web Scraping
- Minimal no of records: 800 only
- 6 kinds of data attributes collected
- Very little minimal Personally Identifiable Information (PII)

```
https://www.youtube.com/watch?v=M73hdSxW1sM
https://www.youtube.com/watch?v=siM4W-4nuMc
https://www.youtube.com/watch?v=ixGaMxHs6A4
https://www.youtube.com/watch?v=kni1w-JxViw
https://www.youtube.com/watch?v=VY6mcZmetYU
https://www.youtube.com/watch?v=LE4SpImAQdY
https://www.youtube.com/watch?v=EtOtuQmaljU
https://www.youtube.com/watch?v=6u2wdQNxs-Q
https://www.youtube.com/watch?v=2rL4foKjtsA
https://www.youtube.com/watch?v=zlmI7ioQ_iY
https://www.youtube.com/watch?v=NatBX3VjnWw
https://www.youtube.com/watch?v=WHLkF566Jn4
https://www.youtube.com/watch?v=MW0UGCIg1tU
https://www.youtube.com/watch?v=MaBiS5JvoYc
https://www.youtube.com/watch?v=CPDxA69zbi8
https://www.youtube.com/watch?v=GZi0Y4HvOS8
https://www.youtube.com/watch?v=NO4ikM9CfYY
https://www.youtube.com/watch?v=Fj6W2iMtKwE
https://www.youtube.com/watch?v=UYvKs8b74qs
https://www.youtube.com/watch?v=ms8wiprrdns
https://www.youtube.com/watch?v=M_VX_8JIe_A
https://www.youtube.com/watch?v=NDYG3j24zko
https://www.youtube.com/watch?v=G5UsHG2HVjM
https://www.youtube.com/watch?v=u9I46dpK_sU
https://www.youtube.com/watch?v=mXNbd9oTvSc
https://www.youtube.com/watch?v=pUibTylbU5k
https://www.youtube.com/watch?v=h45laNr9EvQ
```

When Scraping Video URLs
- Extracted Video URLs only
- Didn’t extract email addresses that may be on the channel pages

Data anonymization for Data crawling
By: Ong Weng Kai

The process of anonymizing data involves altering identifiable information so that it can no longer be traced back to a specific individual. While there are various methods to achieve this, three primary techniques are commonly used:

1. **Suppression**: This basic method removes certain identifying details from the dataset to lessen its traceability.
2. **Generalization**: This technique broadens specific identifiers to less precise categories, such as converting an exact age into a broader age range (e.g., changing 18 to 18-24).
3. **Noise Addition**: This involves swapping identifying data points within a dataset with those from other individuals in the same dataset.(eg. Exchange the zip code between Individual A and Individual B)

However, these methods do not guarantee complete anonymization and must be carefully managed to maintain data utility without compromising privacy.

Data anonymization for Data crawling

In General Data Protection Regulation (GDPR)

three specific reidentification risks:

- **Singling out** — the ability to locate an individual's record within a data set.
- **Linkability** — the ability to link two records pertaining to the same individual or group of individuals.
- **Inference** — the ability to confidently guess or estimate values using other information.

Data anonymization for Data crawling

- In the context of the U.S.,(HIPAA) specifies that data is considered anonymized if 18 specific identifiers are removed, making it no longer "protected".This allows for the data to be used without the same restrictions as identifiable information.

- Names
- Geographic subdivisions smaller than a state
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Fax numbers
- Email addresses
- Social security numbers
- Medical record numbers
- Health insurance beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, such as fingerprints and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

How I applied Data Anonymization

title	post_score	post_id	subreddit	url	comment	comment_score	comment_id	comment_created	reply_number
Call of Duty: Mob...	7	1be6f25	callofdutymobile	https://www.reddi...	Please report any...	1	kvi01uq	1.710803096E9	75
Call of Duty: Mob...	7	1be6f25	callofdutymobile	https://www.reddi...	Hey devs, the new...	27	kurfpzv	1.710375314E9	75

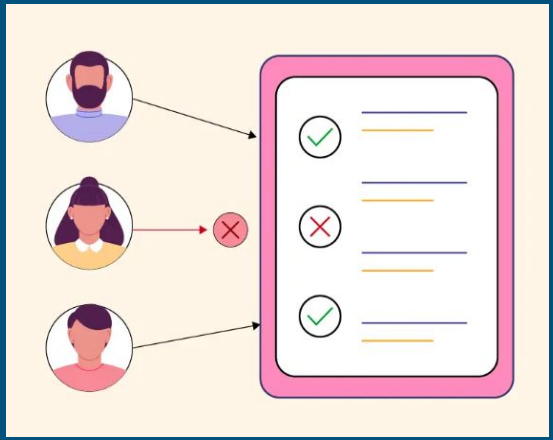
comment
Please report any...
Hey devs, the new...

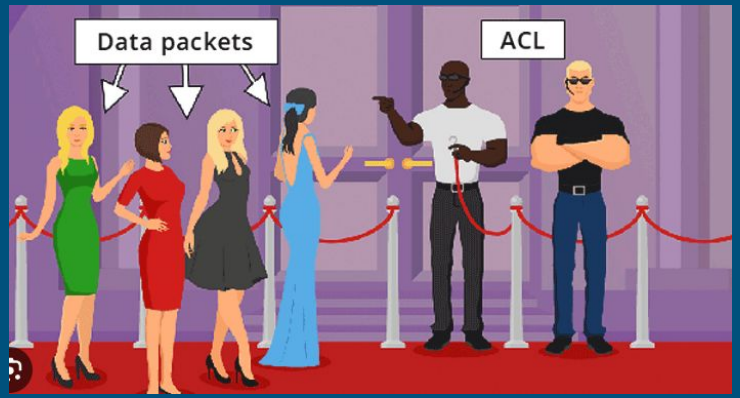
- **Data Anonymization Process:** Involves altering personal data to prevent association with identifiable individuals, ensuring compliance with privacy laws and safeguarding personal information.
- **Removal of Identifiers:** Eliminates specific identifiers such as comment IDs, URLs, and other traceable metadata, retaining only the text of comments.
- **Reduction of Re-identification Risk:** By removing these identifiers, the risk of individuals being traced or identified from the data is significantly reduced.

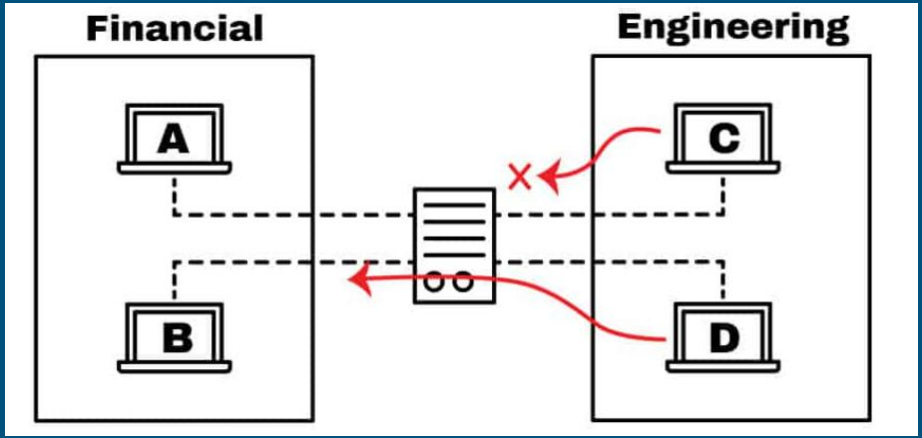
modified_comment	JoinedComment	SentimentScore	Sentiment
Please report any...	please report sea...	0.3182	Positive
Hey devs, the new...	hey devs, new sea...	0.9981	Positive

The Necessity of ACL
By: Sim Hong Li

1. **Data Protection and Privacy**
2. **Regulatory Compliance**
3. **Mitigating Security Risks**







Access Control List Example

Here's a list of some common settings, numerical values and their meanings:

- `-rw-------` (600) -- Only the user has read and write permissions.
- `-rw-r--r--` (644) -- Only user has read and write permissions; the group and others can read only.
- `-rwx------` (700) -- Only the user has read, write and execute permissions.
- `-rwxr-xr-x` (755) -- The user has read, write and execute permissions; the group and others can only read and execute.
- `-rwx--x--x` (711) -- The user has read, write and execute permissions; the group and others can only execute.
- `-rw-rw-rw-` (666) -- Everyone can read and write to the file. Bad idea.
- `-rwxrwxrwx` (777) -- Everyone can read, write and execute. Another bad idea.

Here are a couple common settings for directories:

- `drwx------` (700) -- Only the user can read, write in this directory.
- `drwxr-xr-x` (755) -- Everyone can read the directory, but its contents can only be changed by the user.

Uses Case

```
[hdfs@node1 ~]$ hdfs dfs -mkdir /acltests
[hdfs@node1 ~]$ hdfs dfs -ls /
Found 8 items
drwxr-xr-x   - hdfs    hdfs          0 2016-06-22 19:28 /acltests
drwxrwxrwx   - yarn    hadoop        0 2016-06-22 09:07 /app-logs
drwxr-xr-x   - hdfs    hdfs          0 2015-10-05 20:38 /apps
drwxr-xr-x   - hdfs    hdfs          0 2015-10-05 20:36 /hdp
drwxr-xr-x   - mapred  hdfs          0 2015-10-05 20:36 /mapred
drwxrwxrwx   - mapred  hadoop        0 2015-10-05 20:37 /mr-history
drwxrwxrwx   - hdfs    hdfs          0 2015-10-05 20:41 /tmp
drwxr-xr-x   - hdfs    hdfs          0 2016-06-22 12:39 /user
[hdfs@node1 ~]$
[hdfs@node1 ~]$
[hdfs@node1 ~]$ hdfs dfs -put /etc/passwd /acltests
[hdfs@node1 ~]$ hdfs dfs -ls /acltests
Found 1 items
-rw-r--r--   3 hdfs hdfs       2132 2016-06-22 19:29 /acltests/passwd
[hdfs@node1 ~]$ hdfs dfs -chmod 640 /acltests/passwd
[hdfs@node1 ~]$ hdfs dfs -ls /acltests
Found 1 items
-rw-r-----   3 hdfs hdfs       2132 2016-06-22 19:29 /acltests/passwd
```

```
[hdfs@node1 ~]$ exit
logout
[root@node1 ~]# hdfs dfs -cat /acltests/passwd
cat: Permission denied: user=root, access=READ, inode="/acltests/passwd":hdfs:hdfs:-rw-r-----
[root@node1 ~]#
```

Aspect of Reliability

1. Thong Cheng How
2. Chong Jing Yung
3. Yong Zee Lin

Incident Recovery Plan
By: Thong Cheng How

Incident Recovery Plan

Policies and processes involved

- Backup code we have done into personal computer
- Backup in cloud platform
- Higher availability (if the jupyter notebook fails)
- Version control (if code has bugs)
- Share to trusted partners, employees

Real life examples

Case #5: Massive data breach by two former employees at Tesla

Affected entity	TESLA
Source	Malicious activity by former employees
Consequences	• Personal information of employees and production secrets leaked • Damage to the company's reputation • Potential data protection regulation fines or lawsuits

Case #3: Intellectual property theft by a malicious insider at Yahoo

Affected entity	yahoo!
Source	Malicious insider activity for personal gain
Consequences	• Valuable source code and strategy information leaked • Potential loss of competitive advantage

Related Works

15 Attachments • Scanned by Gmail

ExtractMongoDB ...

PySpark MongoD...

StructuredStrea...

ModellingV5.ipynb

CombinedData (1...

consumer (1).py

HiveDB (1).ipynb

producer (1).py

Risk Assessment for Spark Session
By Chong Jing Yung

Risk Assessment

- Perform regular risk assessments to identify vulnerabilities, threats, and potential impacts on data privacy and security.

In data engineering project:

Avoid Resource Overutilization

- Spark jobs consuming excessive compute resources leading to performance degradation or cluster instability.

Cause

- Run 2 spark session in a same notebook file. (spark and hive_spark)

- No include spark.stop() in notebook.

```
#Hive DB
import findspark
findspark.init()

import pyspark
from pyspark.sql import SparkSession
from pyspark.conf import SparkConf

spark_hive = SparkSession\
         .builder\
         .appName("SparkHiveDemo")\
         .config('spark.sql.warehouse.dir', 'hdfs:/user/hive/warehouse/')\
         .config("spark.sql.catalogImplementation", "hive")\
         .enableHiveSupport()\
         .getOrCreate()
```

```
24/04/20 13:29:41 WARN Utils: Your hostname, jupy-06 resolves to a loopback address: 127.0.1.1; using 10.123.51.206 instead
(on interface ens18)
24/04/20 13:29:41 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/04/20 13:29:43 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
```

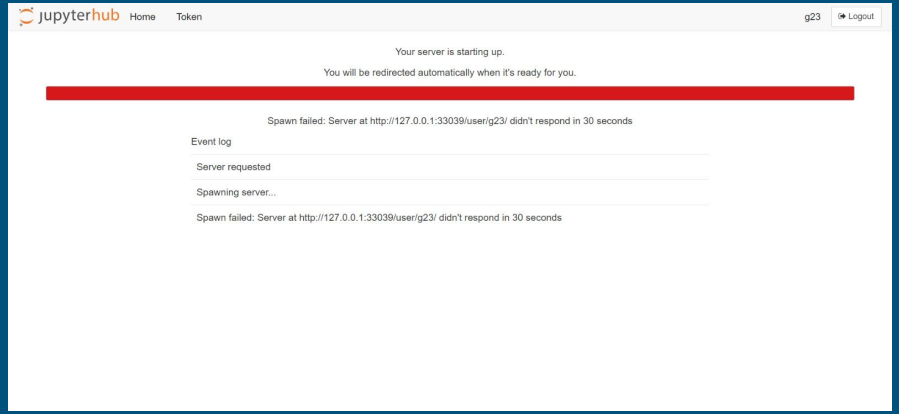
```
#Neo4j
#Extract data from Neo4j to pyspark dataframe for modelling
from pyspark.sql import SparkSession
from neo4j import GraphDatabase

# Create a SparkSession
spark = SparkSession.builder \
    .appName("Neo4j to DataFrame") \
    .getOrCreate()

# Define a function to read data from Neo4j into a PySpark DataFrame
    # Establish connection to Neo4j
graph = GraphDatabase.driver(uri="neo4j+s://46bd93aa.databases.neo4j.io", auth=("neo4j", "M2REJH20Eeod7AZ7n_lLZR4UjWLVacefdrv
session = graph.session()

    # Define Cypher query to retrieve data
cypher_query = "MATCH (c:Combined) RETURN c.comment AS Comment, c.sentimentScore AS SentimentScore, c.sentiment AS Sentiment"

    # Execute Cypher query and retrieve results
result = session.run(cypher_query)

    # Store the results in a list of dictionaries
data = [{"Comment": record["Comment"], "SentimentScore": record["SentimentScore"], "Sentiment": record["Sentiment"]} for reco

    # Close the session
session.close()
```

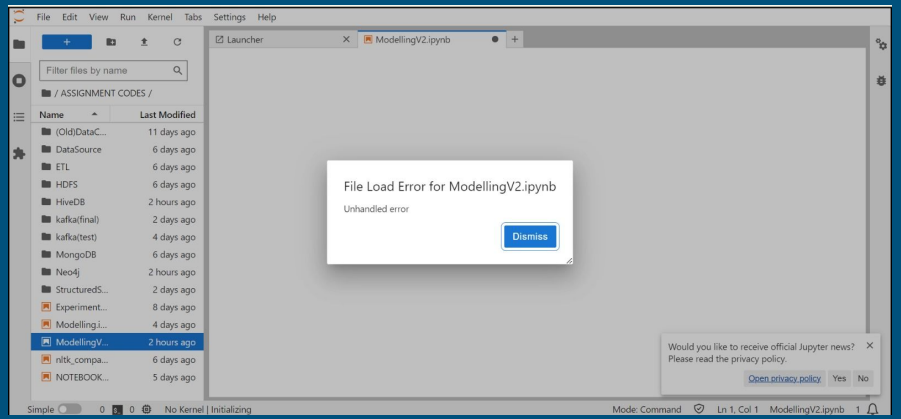
```
24/04/20 13:29:53 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be
loaded.
24/04/20 13:29:55 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
```

Impact

- Performance degradation: When Spark jobs are overly resource-intensive, they take longer to complete and use up less cluster throughput overall.

- Cluster Instability: When a cluster experiences high resource use, it may become unstable and experience job timeouts, cluster crashes, or task failures.





Related Works (Solution)

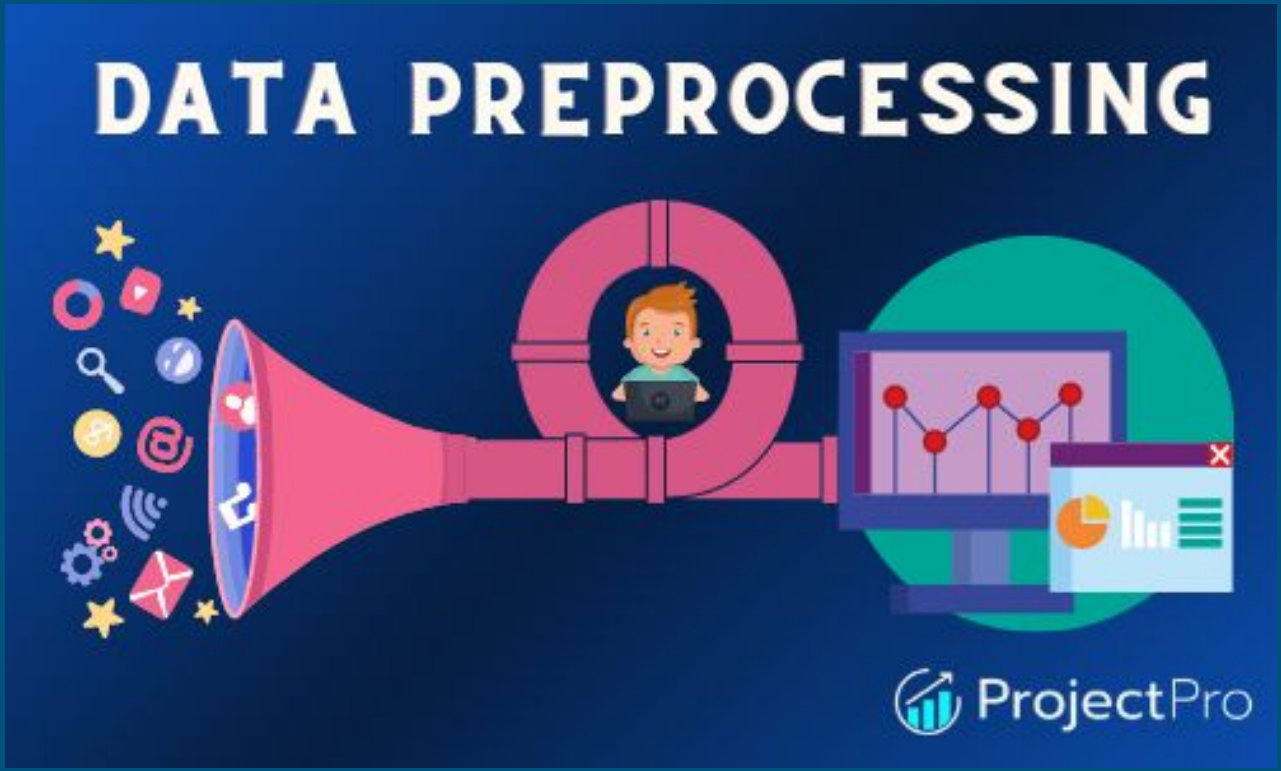
- Use a spark session to handle multitask.
- Stop the spark in the end.

```
spark.stop()
```

```
#Neo4j
#Extract data from Neo4j to pyspark dataframe for modelling
#Hive DB
import findspark
findspark.init()
from pyspark.sql import SparkSession
from pyspark.conf import SparkConf
from neo4j import GraphDatabase

# Create a SparkSession
spark = SparkSession.builder \
    .appName("Modeling") \
    .config('spark.sql.warehouse.dir', 'hdfs:/user/hive/warehouse/')\
    .config("spark.sql.catalogImplementation", "hive")\
    .enableHiveSupport()\
    .getOrCreate()

# Define a function to read data from Neo4j into a PySpark DataFrame
    # Establish connection to Neo4j
graph = GraphDatabase.driver(uri="neo4j+s://46bd93aa.databases.neo4j.io", auth=("neo4j", "M2REJH20Eeod7AZ7n_1LZR4UjWLVacefdrv
session = graph.session()

    # Define Cypher query to retrieve data
cypher_query = "MATCH (c:Combined) RETURN c.comment AS Comment, c.sentimentScore AS SentimentScore, c.sentiment AS Sentiment"

    # Execute Cypher query and retrieve results
result = session.run(cypher_query)

    # Store the results in a list of dictionaries
data = [{"Comment": record["Comment"], "SentimentScore": record["SentimentScore"], "Sentiment": record["Sentiment"]} for reco

    # Close the session
session.close()
```

```
24/04/27 05:03:32 WARN Utils: Your hostname, jupy-06 resolves to a loopback address: 127.0.1.1; using 10.123.51.206 instead
(on interface ens18)
24/04/27 05:03:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/04/27 05:03:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
24/04/27 05:03:33 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
```

Enhance Data Quality Through Preprocessing
By Yong Zee Lin



Why need Data Preprocessing?

- Data preprocessing ensures that the data fed into the model is clean and reliable.
- Without preprocessing, flaws in the data can lead to inaccurate and unreliable results.
- Preprocessing maximizes the accuracy and validity of the model's outcomes.
- In essence, data preprocessing is crucial for obtaining meaningful and actionable insights from your data.

Tokenization, Filtering, and Lemmatization

- **Tokenization:** Breaking down the reviews into individual words or tokens.
- **Filtering:** Filtering out certain words or tokens, likely removing stopwords or irrelevant terms.
- **Lemmatization:** Reducing words to their base or root form, which aids in standardizing variations of words.

reviews	tokenized_reviews	filtered_reviews	lemmatized_reviews
reminds me of the...	[reminds, me, of,...	[reminds, old, co...	[reminds, old, co...
i really really l...	[i, really, reall...	[really, really, ...	[really, really, ...
i love this app s...	[i, love, this, a...	[love, app, somet...	[love, app, somet...
632022 update gla...	[632022, update, ...	[632022, update, ...	[632022, update, ...
freaking phenomen...	[freaking, phenom...	[freaking, phenom...	[freaking, phenom...
exciting game exc...	[exciting, game, ...	[exciting, game, ...	[exciting, game, ...
excellent game wi...	[excellent, game,...	[excellent, game,...	[excellent, game,...
the bots they are...	[the, bots, they,...	[bots, annoying, ...	[bot, annoying, a...
great game but th...	[great, game, but...	[great, game, las...	[great, game, las...
i really like it ...	[i, really, like,...	[really, like, gr...	[really, like, gr...
best fps experien...	[best, fps, exper...	[best, fps, exper...	[best, fps, exper...
its cool game i v...	[its, cool, game,...	[cool, game, play...	[cool, game, play...
great game i love...	[great, game, i, ...	[great, game, lov...	[great, game, lov...
the game is amazi...	[the, game, is, a...	[game, amazing, i...	[game, amazing, i...
the games awesome...	[the, games, awes...	[games, awesome, ...	[game, awesome, i...
okay so the game ...	[okay, so, the, g...	[okay, game, awes...	[okay, game, awes...
the games quite n...	[the, games, quit...	[games, quite, ni...	[game, quite, nic...
great game but th...	[great, game, but...	[great, game, ran...	[great, game, ran...
i find this game ...	[i, find, this, g...	[find, game, fun,...	[find, game, fun,...
simply great cod ...	[simply, great, c...	[simply, great, c...	[simply, great, c...

only showing top 20 rows

Sentiment Analysis

TF-IDF Vectorization:

- Transforms textual data into a numerical format compatible with ML algorithms.

Sentiment Analysis for Tokenized Reviews:

- Analysis of tokenized reviews to determine sentiment polarity.
- Transformation of raw text data into numerical scores, indicating sentiment.
- Represents a form of feature engineering or data transformation in preprocessing textual data.

review	ratings	Sentiment for ratings	SentimentScore	Sentiment for SentimentScore
reminds old cod o...	5	Positive	0.2944	Positive
really really lov...	5	Positive	0.9227	Positive
love app sometime...	5	Positive	0.9524	Positive
632022 update gla...	4	Positive	0.9559	Positive
freaking phenomen...	5	Positive	0.9628	Positive
exciting game exc...	5	Positive	0.9251	Positive
excellent game go...	4	Positive	0.7579	Positive
bot annoying af r...	3	Neutral	-0.7717	Negative
great game last u...	4	Positive	0.5267	Positive
really like graph...	5	Positive	0.9755	Positive
best fps experien...	5	Positive	0.9746	Positive
cool game playing...	4	Positive	-0.4472	Negative
great game love t...	4	Positive	-0.6929	Negative
game amazing imme...	5	Positive	0.34	Positive
game awesome im d...	5	Positive	0.9735	Positive
okay game awesome...	5	Positive	0.93	Positive
game quite nice o...	3	Neutral	-0.8958	Negative
great game random...	5	Positive	0.8126	Positive
find game fun pla...	4	Positive	0.9747	Positive
simply great cod ...	3	Neutral	0.9118	Positive

only showing top 20 rows

Examples of Before and After

Before

_id	ratings	reviews
{660ed074fbd9bf18...	5	Reminds me of the...
{660ed074fbd9bf18...	5	I really, really ...
{660ed074fbd9bf18...	5	I love this app, ...
{660ed074fbd9bf18...	4	6/3/2022 update. ...
{660ed074fbd9bf18...	5	Freaking phenomen...
{660ed074fbd9bf18...	5	Exciting game, ex...
{660ed074fbd9bf18...	4	Excellent game wi...
{660ed074fbd9bf18...	3	The bots.. They a...
{660ed074fbd9bf18...	4	Great game but th...
{660ed074fbd9bf18...	5	I really like it....
{660ed074fbd9bf18...	5	Best fps experien...
{660ed074fbd9bf18...	4	It's cool game I ...
{660ed074fbd9bf18...	4	Great game, i lov...
{660ed074fbd9bf18...	5	The game is amazi...
{660ed074fbd9bf18...	5	The game's awesom...
{660ed074fbd9bf18...	5	Okay so the game ...
{660ed074fbd9bf18...	3	the game's quite ...
{660ed074fbd9bf18...	5	Great game, but t...
{660ed074fbd9bf18...	4	I find this game ...
{660ed074fbd9bf18...	3	Simply great cod ...

only showing top 20 rows

- Preprocessing makes data better for analysis.
- Steps like tokenization, filtering, and lemmatization clean the data.

After

review	ratings	Sentiment for ratings	SentimentScore	Sentiment for SentimentScore
reminds old cod o...	5	Positive	0.2944	Positive
really really lov...	5	Positive	0.9227	Positive
love app sometime...	5	Positive	0.9524	Positive
632022 update gla...	4	Positive	0.9559	Positive
freaking phenomen...	5	Positive	0.9628	Positive
exciting game exc...	5	Positive	0.9251	Positive
excellent game go...	4	Positive	0.7579	Positive
bot annoying af r...	3	Neutral	-0.7717	Negative
great game last u...	4	Positive	0.5267	Positive
really like graph...	5	Positive	0.9755	Positive
best fps experien...	5	Positive	0.9746	Positive
cool game playing...	4	Positive	-0.4472	Negative
great game love t...	4	Positive	-0.6929	Negative
game amazing imme...	5	Positive	0.34	Positive
game awesome im d...	5	Positive	0.9735	Positive
okay game awesome...	5	Positive	0.93	Positive
game quite nice o...	3	Neutral	-0.8958	Negative
great game random...	5	Positive	0.8126	Positive
find game fun pla...	4	Positive	0.9747	Positive
simply great cod ...	3	Neutral	0.9118	Positive

only showing top 20 rows

- TF-IDF turns words into numbers for computers.
- Good preprocessing helps us understand data better and make smarter decisions.

References

- Akhtar, M. M. (2019). Sentiment Analysis on Youtube Comments: A brief study. Data Mining Project, 3–3. https://www.researchgate.net/publication/344013948_Sentiment_Analysis_on_Youtube_Comments_A_brief_study
- Bachubhay, A., Chhour, D., Deng, H., & Tran, T. (2021). YouTube Video Analysis. CS4624 (Multimedia, Hypertext, and Information Access). https://doi.org/https://vtechworks.lib.vt.edu/server/api/core/bitstreams/c831882a-0db8-46eb-9c83-35ef9e8d5cb0/content
- EDPS. (2024, May 3). D. European Data Protection Supervisor. https://www.edps.europa.eu/data-protection/data-protection/glossary/d_en#:~:text=The%20principle%20of%20%E2%80%9Cdata%20minimisation,necessary%20to%20fulfil%20that%20purpose.
- G, A., & P, R. S. (2022). Sentimental Analysis on YouTube scrapped data. International Journal of Advanced Research in Science, Communication and Technology, 2(6), 667–671. https://doi.org/10.48175/ijarsct-5090
- Oh, P. (2023, October 14). Balancing act: Choosing between data minimization and data retention for PDPA compliance. Medium. https://medium.com/datafrens-sg/balancing-act-choosing-between-data-minimization-and-data-retention-for-pdpa-compliance-1b6bf0e44491
- *Premium vector: Disaster recovery infographic template design with icons Vector Illustration Technology Concept*. Freepik. (2024). https://www.freepik.com/premium-vector/disaster-recovery-infographic-template-design-with-icons-vector-illustration-technology-concept_31955453.htm
- Pryimenko, L. (2024). *7 real-life data breaches caused by insider threats | ekran system*. 7 examples of Real-life data breaches caused by insider threats. https://www.ekransystem.com/en/blog/real-life-examples-insider-threat-caused-breaches
- WinZip. (2023). Articles. WinZip Enterprise Blog. https://winzip.com/blog/enterprise/understanding-data-minimization/?alid=245629300.1714343722


THANK YOU