databricks 2024-02-25 - DBFS Example

(https://databricks.com)

Overview

This notebook will show you how to create and query a table or DataFrame that you uploaded to DBFS. DBFS (https://docs.databricks.com/user-guide/dbfs-databricks-file-system.html) is a Databricks File System that allows you to store data for querying inside of Databricks. This notebook assumes that you have a file already inside of DBFS that you would like to read from.

This notebook is written in **Python** so the default cell type is Python. However, you can use different languages by using the `%LANGUAGE` syntax. Python, Scala, SQL, and R are all supported.

```
# File location and type
file_location = "/FileStore/tables/users-1.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df = spark.read.format(file_type) \
  .option("inferSchema", infer_schema) \
  .option("header", first_row_is_header) \
  .option("sep", delimiter) \
  .load(file_location)

display(df)
```

Table

	id	user	age
1	0	_TheSpecialOne_	unknown
2	303	scotthamilton	unknown
3	548	mattycus	unknown
4	815	ElleCTF	unknown
5	824	Karoli	unknown
6	1003	joy_wolf	unknown
7	1223	mybirch	unknown

10,000 rows | Truncated data

```
# File location and type
file_location = "/FileStore/tables/followers_new.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df1 = spark.read.format(file_type) \
  .option("inferSchema", infer_schema) \
  .option("header", first_row_is_header) \
  .option("sep", delimiter) \
  .load(file_location)


display(df1)
```

Table

	id	following
1	210499023.0	587419676.0
2	576194305.0	585452007.0
3	603493582.0	591848745.0
4	763712420.0	521784660.0
5	209141897.0	85858045.0
6	595342589.0	595353768.0
7	784146077.0	326516445.0

10,000 rows | Truncated data

```
# File location and type
file_location = "/FileStore/tables/tweets_new.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df2 = spark.read.format(file_type) \
  .option("inferSchema", infer_schema) \
  .option("header", first_row_is_header) \
  .option("sep", delimiter) \
  .load(file_location)


display(df2)
```

Table

	id	timestamp	text
1	0	2009-04-06 22:19:45	@switchfoot http://twitpic.com/2y1zl - awww, that's a bummer. you shoulda got david carr of third day to do it
2	303	2009-04-06 22:19:49	is upset that he can't update his facebook by texting it... and might cry as a result school today also. blah!
3	548	2009-04-06 22:19:53	@kenichan i dived many times for the ball. managed to save 50% the rest go out of bounds
4	815	2009-04-06 22:19:57	my whole body feels itchy and like its on fire
5	824	2009-04-06 22:19:57	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because i can't see you all over there.
6	1003	2009-04-06 22:20:00	@kwesidei not the whole crew
7	1223	2009-04-06 22:20:03	need a hug

10,000 rows | Truncated data

```
# Create a view or table

temp_table_name = "users"

df.createOrReplaceTempView(temp_table_name)

temp_table_name = "followers"

df1.createOrReplaceTempView(temp_table_name)

temp_table_name = "tweets"

df2.createOrReplaceTempView(temp_table_name)
```

Tweets count by age group

```
%sql
/* tweet count by age group*/
SELECT age,
       COUNT(*) AS tweet_count
FROM users u
JOIN tweets t ON u.id = t.id
GROUP BY age;
```

Table

	age	tweet_count
1	unknown	1568803
2	old	13214
3	young	18003

3 rows

Insights:This table shows the highest number of tweets for the unknown category. Younger generation has the highest number of tweets which is 18003 and older generation has 13214 tweets.

which age groups engage the most with tweets containing specific texts

```
%sql
/*which age groups engage the most with tweets containing specific texts*/
SELECT u.age,
       t.text,
       COUNT(*) AS hashtag_count
FROM users u
JOIN tweets t ON u.id = t.id
WHERE t.text IS NOT NULL
GROUP BY u.age, t.text
ORDER BY u.age, hashtag_count DESC;
```

Table

	age	text	hashtag
1	old	@estrogen: "@blippo thanx! a long family day *g*" cool y shouldn't birds sing!! http://tinyurl.com/cco-glo... http://blip.fm/~7sfbm	6
2	old	missing my family	6

3	old	painting	6
4	old	painting my nails	6
5	old	@itsjoejonas ohhhh is there anyway u guys can come tour here? i would travel the world just to c u guys perform, please try and come xoxo	5
6	old	watching family guy	5
7	old	freebie: bass pro shops are holding free family classes all summer! http://short.to/1cp2 follow me for your twitter freebies!	4

10,000 rows | Truncated data

Insights: This table that indicates a pattern of older generations using the word "family" more frequently in their tweets compared to younger generations. Older generations may prioritize family and traditional values more prominently in their lives, which reflects in their online communication. Older individuals may be more likely to have families of their own or be more involved in family activities, leading to the increased mention of family-related topics in their tweets.

To find tweets related to business, politics, and education

```
%sql
/* I want to find specific keyword in tweets*/
SELECT keyword,
       COUNT(*) AS keyword_count
FROM
  (SELECT id,
          text,
          'keyword1' AS keyword
   FROM tweets
   WHERE text LIKE '%business%'
   UNION ALL
   SELECT id,
          text,
          'keyword2' AS keyword
   FROM tweets
   WHERE text LIKE '%politics%'
   UNION ALL
   SELECT id,
          text,
          'keyword3' AS keyword
   FROM tweets
   WHERE text LIKE '%education%') AS keyword_tweets
GROUP BY keyword;
```

Table

	keyword	keyword_count
1	keyword1	2273
2	keyword2	299
3	keyword3	276

3 rows

Insights:The word "business" was used 2,273 times. The word "politics" was used 299 times. The word "education" was used 276 times. These numbers give you an indication of the relative frequency of these terms in the tweets you analyzed. It suggests that "business" is the most commonly mentioned topic, followed by "politics" and then "education."

The word "business" within a specific date range

```
%sql

/*SELECT text,timestamp FROM tweets WHERE  text like '%business%' and timestamp "2009-04-06" and "2009-04-13";*/
SELECT text,timestamp
FROM tweets
WHERE text LIKE '%business%' and timestamp BETWEEN "2009-04-06" AND "2009-04-13"
```

Table

	text	timestamp
1	gym attire today was: puma singlet, adidas shorts.....and black business socks and leather shoes lucky did not run into any cute girls.	2009-04-06 22:23:43
2	bad news was dad has cancer and is dying good news new business started and i am now a life coach practising holistic weight management	2009-04-06 22:45:40
3	why do other pet care people try to run others out of business? or send suspicious e-mails fishing for info?	2009-04-06 23:18:32
4	doing some business studies revision	2009-04-06 23:55:40
5	school work boring, working on a business law paper right now can't wait till i'm done. hopefully tomorrow grrrr...	2009-04-07 00:46:23
6	@mathie makes for hard work for small businesses	2009-04-07 01:27:27
7	still in the business plan meeting.. too many back to back presentations	2009-04-07 02:02:45

39 rows

Insights:There were 39 tweets containing the word "business" between April 6, 2009, and April 13, 2009, it indicates the frequency of tweets mentioning "business" during that time frame. This information can provide insights into the discussions or trends related to business during that particular week in 2009. It might suggest a heightened interest or activity in business-related topics during that period.

Identify the time

```
%sql
/*to identify the time period during which most tweets were posted*/

SELECT HOUR(timestamp) AS tweet_hour,
       COUNT(*) AS tweet_count
FROM tweets
GROUP BY tweet_hour
ORDER BY tweet_count DESC
LIMIT 5;
```

Table

	tweet_hour	tweet_count
1	23	84750
2	7	83654
3	0	80865
4	6	80852
5	5	78623

5 rows

Insights: The highest tweet volume occurred at hour 23 with 84,750 tweets, and the second highest occurred at hour 7 with 83,654 tweets, it suggests that these hours were particularly active on Twitter. Analyzing tweet volumes by hour can reveal patterns in user activity, which could be influenced by various factors such as time zones, global events, or online trends. This data could be useful for

understanding peak hours of engagement on the platform or for planning social media strategies to reach a larger audience.

Age group and time period

```
%sql
/*To identify the time period during which most tweets were posted and the corresponding age group of the users*/

SELECT u.age AS user_age,
       HOUR(timestamp) AS tweet_hour
FROM tweets t
JOIN users u ON t.id = u.id

ORDER BY tweet_hour DESC
LIMIT 30
```

Table

	user_age	tweet_hour
1	unknown	23
2	unknown	23
3	unknown	23
4	unknown	23
5	unknown	23
6	unknown	23
7	unknown	23

30 rows

Insights:This table shows that the younger generation tweets more than the older generation in 23 hours. The data indicates that the younger generation consistently tweets more than the older generation in each of these hours, it suggests a trend of higher Twitter activity among younger individuals throughout the day.

Hashtag

```
%sql
/*To identify trending hashtag in tweets between 2009-04-06 and 2009-04-20*/
SELECT SUBSTRING_INDEX(SUBSTRING_INDEX(text, '#', numbers.n), ' ', -1) AS hashtag,
       COUNT(*) AS hashtag_count
FROM
  (SELECT 1 AS n UNION ALL SELECT 2 UNION ALL SELECT 3 UNION ALL SELECT 4 UNION ALL SELECT 5) numbers
JOIN tweets ON CHAR_LENGTH(text) - CHAR_LENGTH(REPLACE(text, '#', '')) >= n - 1
WHERE timestamp >= '2009-04-06' AND timestamp <= '2009-04-20'
GROUP BY hashtag
ORDER BY hashtag_count DESC
LIMIT 10;
```

Table

	hashtag	hashtag_count
1		56463
2	x	503
3	lol	449
4	xx	248
5	#asot400	240
6	<3	225
7	;)	169

10 rows

Insights:The 'X' shows 503 times in tweets during the time period between 2009/04/06 and 2009/04/20.

Tweets containing links

Table

	id	text	timesta
1	0	@switchfoot http://twitpic.com/2y1zl - awww, that's a bummer. you shoulda got david carr of third day to do it. ;d	2009-04
2	7005	@missxu sorry! bed time came here (gmt+1) http://is.gd/fnge	2009-04
3	10494	broadband plan 'a massive broken promise' http://tinyurl.com/dcuc33 via www.diigo.com/~tautao still waiting for broadband we are	2009-04
4	15683	why won't you show my location?! http://twitpic.com/2y2es	2009-04
5	26079	strider is a sick little puppy http://apps.facebook.com/dogbook/profile/view/5248435	2009-04
6	26504	body of missing northern calif. girl found: police have found the remains of a missing northern california girl .. http://tr.im/imji	2009-04
7	27393	emily will be glad when mommy is done training at her new job. she misses her. http://apps.facebook.com/dogbook/profile/view/6176014	2009-04

10,000 rows | Truncated data