# Final Project M3

CS 327E
November 13, 2017

1) What is the core abstraction in Spark SQL?

A) MapReduce
B) RDDs
C) DataFrames
D) Catalyst

2) Like RDDs, DataFrames are immutable.

A) True
B) False

3) Unlike RDDs, DataFrames require _____

A) Schemas
B) Relational Databases
C) User Defined Functions
D) User Defined Types

4) How can you explore the data that is in a DataFrame?

A) With SQL expressions
B) Using the DataFrame API
C) By converting the DataFrame to an RDD and then running RDD operations over the RDD
D) All of the above

# 5) How many rows are returned by `distinct_table.show()`?

```
1  sc = SparkContext()
2  sqlctx = SQLContext(sc)
3
4  def create_rows(tuple):
5    id, tag = tuple
6    row = Row(id=id, tag=tag)
7    return row
8
9  base_rdd = sc.parallelize((1, "time travel"), (1, "adventure"), (2, "travel"), (2, "travel")]
10 row_rdd = base_rdd.map(create_rows)
11
12 title_tag_table = sqlctx.createDataFrame(row_rdd)
13 title_tag_table.createOrReplaceTempView("Tags")
14
15 distinct_table = sqlctx.sql("SELECT DISTINCT id, tag FROM Tags")
16 print distinct_table.show()
```

A) 2 B) 3 C) 4

# Querying Postgres from Python

```
13  conn = psycopg2.connect(database=rds_database, user=rds_user, password=rds_password, host=rds_host, port=rds_port)
14
15  select_stmt = "select count(*) from title_basics where start_year = 2017"
16
17  try:
18
19      cur = conn.cursor()
20      cur.execute(select_stmt)
21
22      row = cur.fetchone()
23
24      if row != None:
25          record_count = row[0]
26          print("record count is: " + str(record_count))
27
28  except Exception as e:
29      print("Exception: " + e.message)
30
31      cur.close()
32      conn.close()
```

# Querying Postgres from Python

```
13  conn = psycopg2.connect(database=rds_database, user=rds_user, password=rds_password, host=rds_host, port=rds_port)
14
15  select_stmt = "select title_id, primary_title from title_basics where start_year = 2017"
16
17  try:
18
19      cur = conn.cursor()
20      cur.execute(select_stmt)
21
22      rows = cur.fetchall()
23
24      for row in rows:
25          title_id = row[0]
26          primary_title = row[1]
27          print("title_id: " + title_id + ", primary_title: " + primary_title)
28
29  except Exception as e:
30      print("Exception: " + e.message)
31
32      cur.close()
33      conn.close()
```

# Milestone 3

http://www.cs.utexas.edu/~scohen/projects/m3-assignment.pdf