Exam

April 27, 2024

1 IST769 Final Exam

INSTRUCTIONS FOR HIGHEST GRADE POSSIBLE

Unless you are explicitly instructed otherwise, answer each of the following using PySpark / Spark SQL. For any queries you write make sure to include a `printSchema()` and sample(s) of the output which clearly demonstrates the code is correct.

[1]:

```
! sudo cp /home/jovyan/work/jars/neo4j-connector-apache-spark_2.12-4.1.
↪0_for_spark_3.jar /usr/local/spark/jars/neo4j-connector-apache-spark_2.12-4.
↪1.0_for_spark_3.jar
```

[2]:

```
!pip install -q cassandra-driver
```

[3]:

```
import pyspark
from pyspark.sql import SparkSession
```

[4]:

```
# YOUR NAME ========> Sravan Kumar Mangalagiri
# YOUR SU EMAIL ====> smangala@syr.edu
```

1.0.1 Question 1

In the cell below configure a spark session that is configured to connect to `mongodb`, `minio`, `cassandra`, ’`elasticsearch` and `neo4j`.

[5]:

```
#1 Spark session

import pyspark
from pyspark.sql import SparkSession

user = "mongo"
passwd = "SU2orange!"
s3_bucket = "gamestreams"
s3_server = "http://minio:9000"
s3_access_key = "minio"
s3_secret_key = "SU2orange!"
elastic_host = "elasticsearch"
elastic_port = "9200"
```

```
cassandra_host = "cassandra"
bolt_url = "bolt://neo4j:7687"

mongo_uri = "mongodb://admin:mongopw@mongo:27017/admin?authSource=admin"

jars = [
    "com.datastax.spark:spark-cassandra-connector-assembly_2.12:3.1.0",
    "org.elasticsearch:elasticsearch-spark-20_2.12:7.15.0",
    "org.mongodb.spark:mongo-spark-connector_2.12:3.0.1"
]

spark = SparkSession.builder \
    .master("local") \
    .appName('jupyter-pyspark') \
        .config("spark.jars.packages",",".join(jars) )\
        .config("spark.hadoop.fs.s3a.endpoint", s3_server ) \
        .config("spark.hadoop.fs.s3a.access.key", s3_access_key) \
        .config("spark.hadoop.fs.s3a.secret.key", s3_secret_key) \
        .config("spark.hadoop.fs.s3a.fast.upload", True) \
        .config("spark.hadoop.fs.s3a.path.style.access", True) \
        .config("spark.hadoop.fs.s3a.impl", "org.apache.hadoop.fs.s3a.
 ↪S3AFileSystem") \
        .config("spark.cassandra.connection.host", cassandra_host) \
        .config("spark.es.nodes", elastic_host) \
        .config("spark.es.port",elastic_port) \
        .config("spark.mongodb.input.uri", mongo_uri) \
        .config("spark.mongodb.output.uri", mongo_uri) \
        .getOrCreate()
sc = spark.sparkContext
sc.setLogLevel("ERROR") # Keeps the noise down!!!
```

```
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/usr/local/spark-3.1.2-bin-hadoop3.2/jars/spark-unsafe_2.12-3.1.2.jar) to
constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of
org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal
reflective access operations
WARNING: All illegal access operations will be denied in a future release
```

```
:: loading settings :: url = jar:file:/usr/local/spark-3.1.2-bin-
hadoop3.2/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
```

```
Ivy Default Cache set to: /home/jovyan/.ivy2/cache
The jars for the packages stored in: /home/jovyan/.ivy2/jars
com.datastax.spark#spark-cassandra-connector-assembly_2.12 added as a dependency
org.elasticsearch#elasticsearch-spark-20_2.12 added as a dependency
org.mongodb.spark#mongo-spark-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-
parent-730830d0-8bfe-49c4-9455-530ef4754903;1.0
        confs: [default]
        found com.datastax.spark#spark-cassandra-connector-assembly_2.12;3.1.0
in central
        found org.elasticsearch#elasticsearch-spark-20_2.12;7.15.0 in central
        found org.scala-lang#scala-reflect;2.12.8 in central
        found org.slf4j#slf4j-api;1.7.6 in central
        found commons-logging#commons-logging;1.1.1 in central
        found javax.xml.bind#jaxb-api;2.3.1 in central
        found com.google.protobuf#protobuf-java;2.5.0 in central
        found org.apache.spark#spark-yarn_2.12;2.4.4 in central
        found org.mongodb.spark#mongo-spark-connector_2.12;3.0.1 in central
        found org.mongodb#mongodb-driver-sync;4.0.5 in central
        found org.mongodb#bson;4.0.5 in central
        found org.mongodb#mongodb-driver-core;4.0.5 in central
downloading https://repo1.maven.org/maven2/com/datastax/spark/spark-cassandra-
connector-assembly_2.12/3.1.0/spark-cassandra-connector-assembly_2.12-3.1.0.jar
…
        [SUCCESSFUL ] com.datastax.spark#spark-cassandra-connector-
assembly_2.12;3.1.0!spark-cassandra-connector-assembly_2.12.jar (543ms)
downloading https://repo1.maven.org/maven2/org/elasticsearch/elasticsearch-
spark-20_2.12/7.15.0/elasticsearch-spark-20_2.12-7.15.0.jar …
        [SUCCESSFUL ] org.elasticsearch#elasticsearch-
spark-20_2.12;7.15.0!elasticsearch-spark-20_2.12.jar (82ms)
downloading https://repo1.maven.org/maven2/org/mongodb/spark/mongo-spark-
connector_2.12/3.0.1/mongo-spark-connector_2.12-3.0.1.jar …
        [SUCCESSFUL ] org.mongodb.spark#mongo-spark-connector_2.12;3.0.1!mongo-
spark-connector_2.12.jar (29ms)
downloading https://repo1.maven.org/maven2/org/mongodb/mongodb-driver-
sync/4.0.5/mongodb-driver-sync-4.0.5.jar …
        [SUCCESSFUL ] org.mongodb#mongodb-driver-sync;4.0.5!mongodb-driver-
sync.jar (25ms)
downloading https://repo1.maven.org/maven2/org/mongodb/bson/4.0.5/bson-4.0.5.jar
…
        [SUCCESSFUL ] org.mongodb#bson;4.0.5!bson.jar (24ms)
downloading https://repo1.maven.org/maven2/org/mongodb/mongodb-driver-
core/4.0.5/mongodb-driver-core-4.0.5.jar …
        [SUCCESSFUL ] org.mongodb#mongodb-driver-core;4.0.5!mongodb-driver-
core.jar (58ms)
:: resolution report :: resolve 6634ms :: artifacts dl 780ms
        :: modules in use:
        com.datastax.spark#spark-cassandra-connector-assembly_2.12;3.1.0 from
central in [default]
        com.google.protobuf#protobuf-java;2.5.0 from central in [default]
        commons-logging#commons-logging;1.1.1 from central in [default]
        javax.xml.bind#jaxb-api;2.3.1 from central in [default]
        org.apache.spark#spark-yarn_2.12;2.4.4 from central in [default]
        org.elasticsearch#elasticsearch-spark-20_2.12;7.15.0 from central in
[default]
        org.mongodb#bson;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-core;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-sync;4.0.5 from central in [default]
        org.mongodb.spark#mongo-spark-connector_2.12;3.0.1 from central in
[default]
        org.scala-lang#scala-reflect;2.12.8 from central in [default]
        org.slf4j#slf4j-api;1.7.6 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   12  |   12  |   12  |   0   ||   6   |   6   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-
parent-730830d0-8bfe-49c4-9455-530ef4754903
        confs: [default]
        6 artifacts copied, 0 already retrieved (19325kB/108ms)
24/04/27 01:34:36 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
```

1.0.2 Question 2

Demonstrate you can read the process-oriented data `enrollments` and `sections` from `minio` using PySpark.

[6]:

```
#2a enrollments
enrollment = spark.read.option("header", True).option("inferSchema", True).
 ↪option("sep", ",").csv("s3a://enrollments/enrollments.csv")\
.toDF("term", "course_enrollment", "course", "section", "student_id", "grade",␣
 ↪"grade_points")

enrollment.show()
enrollment.printSchema()
enrollment.count()
```

```
+----+-----------------+------+-------+-----------------+-----+------------+
|term|course_enrollment|course|section|       student_id|grade|grade_points|
+----+-----------------+------+-------+-----------------+-----+------------+
|1221|                1|IST659|   M001|      orenjouglad|    C|         2.0|
|1221|                2|IST659|   M001|      billmelator|    A|         4.0|
|1221|                3|IST659|   M001|       morrisless|    A|         4.0|
|1221|                4|IST659|   M001|amberwavesofgrain|   A-|       3.667|
|1221|                5|IST659|   M001|         abbykuss|    A|         4.0|
|1221|                6|IST659|   M001|       tallyitupp|    A|         4.0|
|1221|                7|IST659|   M001|     rubyslippers|   B-|       2.667|
|1221|                8|IST659|   M001|          salladd|   A-|       3.667|
|1221|                9|IST659|   M001|isabellegunnering|    A|         4.0|
|1221|               10|IST659|   M001|        rustycarz|    B|         3.0|
|1221|               11|IST659|   M001|      pattyo'beef|    A|         4.0|
|1221|               12|IST659|   M001|        tyitdowne|    A|         4.0|
|1221|               13|IST659|   M001|        windysees|    A|         4.0|
|1221|               14|IST659|   M001|      gingersnaps|   A-|       3.667|
|1221|               15|IST659|   M001|        harrypits|    A|         4.0|
|1221|               16|IST659|   M001|        frankklee|    A|         4.0|
|1221|               17|IST659|   M001|   chrispeanugget|    A|         4.0|
|1221|               18|IST659|   M001|    isrealornotte|   A-|       3.667|
|1221|               19|IST659|   M001|      windyshores|    A|         4.0|
|1221|               20|IST659|   M001|     mistymeadows|    A|         4.0|
+----+-----------------+------+-------+-----------------+-----+------------+
only showing top 20 rows

root
 |-- term: integer (nullable = true)
 |-- course_enrollment: integer (nullable = true)
 |-- course: string (nullable = true)
 |-- section: string (nullable = true)
 |-- student_id: string (nullable = true)
 |-- grade: string (nullable = true)
 |-- grade_points: double (nullable = true)
```

[6]: 743

[7]:

```
#2b sections
sections = spark.read.option("header", True).option("inferSchema", True).
 ↪option("sep", ",").csv("s3a://enrollments/sections.csv")\
.toDF("term", "course","section","enrollment", "capacity")

sections.show()
sections.printSchema()
```

```
+----+------+-------+----------+--------+
|term|course|section|enrollment|capacity|
+----+------+-------+----------+--------+
|1221|IST659|   M001|        20|      20|
|1221|IST659|   M002|        20|      20|
|1221|IST722|   M001|        25|      28|
|1221|IST615|   M001|        22|      28|
|1221|IST621|   M001|        22|      24|
|1221|IST687|   M001|        20|      20|
|1221|IST687|   M002|        21|      24|
|1221|IST707|   M001|        28|      28|
|1222|IST659|   M001|        24|      24|
|1222|IST769|   M001|        19|      24|
|1222|IST615|   M001|        19|      24|
|1222|IST714|   M001|        17|      20|
|1222|IST621|   M001|        28|      28|
|1222|IST621|   M002|        22|      24|
|1222|IST687|   M001|        18|      20|
|1222|IST687|   M002|        20|      20|
|1222|IST718|   M001|        28|      28|
|1231|IST659|   M001|        20|      20|
|1231|IST659|   M002|        20|      20|
|1231|IST722|   M001|        23|      28|
+----+------+-------+----------+--------+
only showing top 20 rows

root
 |-- term: integer (nullable = true)
 |-- course: string (nullable = true)
 |-- section: string (nullable = true)
 |-- enrollment: integer (nullable = true)
 |-- capacity: integer (nullable = true)
```

1.0.3 Question 3

Demonstrate you can read the reference-oriented data `terms`, `students`, `courses`, and `program` reference data from `MongoDb` using PySpark.

[8]:
```
#3a terms
terms = spark.read.format("mongo").option("database","ischooldb").
↪option("collection","terms").load()
terms.show()
terms.printSchema()
```

```
+----+-------------+----+-----------+--------+----+
| _id|academic_year|code|       name|semester|year|
+----+-------------+----+-----------+--------+----+
|1221|    2021-2022|1221|  Fall 2021|    Fall|2021|
|1222|    2021-2022|1222|Spring 2022|  Spring|2022|
|1231|    2022-2023|1231|  Fall 2022|    Fall|2022|
|1232|    2022-2023|1232|Spring 2023|  Spring|2023|
+----+-------------+----+-----------+--------+----+

root
```

```
 |-- _id: string (nullable = true)
 |-- academic_year: string (nullable = true)
 |-- code: string (nullable = true)
 |-- name: string (nullable = true)
 |-- semester: string (nullable = true)
 |-- year: integer (nullable = true)
```

[9]:

```
#3b courses
courses = spark.read.format("mongo").option("database","ischooldb").
 ↪option("collection","courses").load()
courses.show(100)
courses.printSchema()
```

```
+------+------+-------+--------------------+--------------------+---------------
-+--------------------+-------------+--------------------+
|   _id|  code|credits|         description|elective_in_programs|
key_assignments|                name|prerequisites|required_in_programs|
+------+------+-------+--------------------+--------------------+---------------
-+--------------------+-------------+--------------------+
|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|IST722|IST722|      3|Introduction to c…|                [IS]| [project,
exam]|    Data Warehousing|     [IST659]|                  []|
|IST769|IST769|      3|Analyze relationa…|                [DS]| [project,
exam]|Advanced Big Data…|     [IST659]|                  []|
|IST615|IST615|      3|Cloud services cr…|                  []|[project,
paper]|    Cloud Management|           []|            [IS, DS]|
|IST714|IST714|      3|Advanced, lab-bas…|            [IS, DS]|
[project]|  Cloud Architecture|     [IST615]|                  []|
|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|IST687|IST687|      3|Introduces inform…|                [IS]| [project,
exam]|Introduction to D…|           []|                [DS]|
|IST707|IST707|      3|General overview …|                [IS]|
[exam]|Applied Machine L…|     [IST687]|                [DS]|
|IST718|IST718|      3|A broad introduct…|                  []|
[project]|  Big Data Analytics|     [IST687]|                [DS]|
+------+------+-------+--------------------+--------------------+---------------
-+--------------------+-------------+--------------------+

root
 |-- _id: string (nullable = true)
 |-- code: string (nullable = true)
 |-- credits: integer (nullable = true)
 |-- description: string (nullable = true)
 |-- elective_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- key_assignments: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- name: string (nullable = true)
 |-- prerequisites: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- required_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
```

[10]:
```
#3c Programs
programs = spark.read.format("mongo").option("database","ischooldb").
 ↪option("collection","programs").load()
programs.show(100)
programs.printSchema()
```

```
+---+----+-------+--------------------+--------------------+--------------------
+-----------+
|_id|code|credits|    elective_courses|                name|
required_courses|       type|
+---+----+-------+--------------------+--------------------+--------------------
+-----------+
| IS|  IS|     36|[IST722, IST714, …| Information Systems|[IST659, IST615,
…|    Masters|
| DS|  DS|     34|    [IST769, IST714]|        Data Science|[IST659, IST615,
…|    Masters|
|BDC| BDC|      9|                null|Data Engineering …|[IST659, IST722,
…|Certificate|
|CCC| CCC|      9|                null|Cloud Computing C…|[IST621, IST615,
…|Certificate|
|MLC| MLC|      9|                null|Machine Learning …|[IST687, IST707,
…|Certificate|
+---+----+-------+--------------------+--------------------+--------------------
+-----------+

root
 |-- _id: string (nullable = true)
 |-- code: string (nullable = true)
 |-- credits: integer (nullable = true)
 |-- elective_courses: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- name: string (nullable = true)
 |-- required_courses: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- type: string (nullable = true)
```

```
[11]: #3d students
students = spark.read.format("mongo").option("database","ischooldb").
 ↪option("collection","students").load()
students.show()
students.printSchema()
students.count()
```

```
+-----------------+------------------+-------+
|              _id|              name|program|
+-----------------+------------------+-------+
|         abbykuss|         Abby Kuss|     DS|
|       adamantium|       Adam Antium|     IS|
|        addieowse|        Addie Owse|     IS|
|     aidensomewun|     Aiden Somewun|     IS|
|     aidenknowone|     Aiden Knowone|     DS|
|         alfrecso|         Al Frecso|     DS|
|          alkohol|          Al Kohol|     DS|
|      allanwrench|      Allan Wrench|     IS|
|        allygator|        Ally Gator|     IS|
|  almafrienzergon|  Alma Frienzergon|     IS|
| amandahugginkiss| Amanda Hugginkiss|     IS|
|amberwavesofgrain|Amber Wavesofgrain|     DS|
|         anitajob|         Anita Job|     IS|
|       anitafavor|       Anita Favor|     IS|
|      anitashower|      Anita Shower|     DS|
|    anitasandwich|    Anita Sandwich|     DS|
|        annedewey|        Anne Dewey|     IS|
|       aprilfirst|       April First|     DS|
|       arialphoto|       Arial Photo|     DS|
| arialsurvellence| Arial Survellence|     IS|
+-----------------+------------------+-------+
only showing top 20 rows

root
 |-- _id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- program: string (nullable = true)
```

```
[11]: 235
```

1.0.4 Question 4

Prepare the section data for loading into cassandra and elasticsearch with Spark or Spark SQL. Just PREPARE it do not LOAD it. Remember that we want this data to be as wide as possible, so include all relevant reference data. For example, the section data should include term attributes like year, academic year, etc… and from course, attributes like credits, name, prerequisites, etc…

[12]:
```
#4 wide_sections
#743
courses = courses.withColumnRenamed("name","course_name")
combined = sections.join(terms,sections["term"] == terms["_id"],"inner")
section = combined.join(courses,combined["course"] == courses["_id"],"inner")
section.count()
section.printSchema()
section.show(34)
section = section.
 ↪select('term','course','section','enrollment','capacity','academic_year','name','semester',
```

```
root
 |-- term: integer (nullable = true)
 |-- course: string (nullable = true)
 |-- section: string (nullable = true)
 |-- enrollment: integer (nullable = true)
 |-- capacity: integer (nullable = true)
 |-- _id: string (nullable = true)
 |-- academic_year: string (nullable = true)
 |-- code: string (nullable = true)
 |-- name: string (nullable = true)
 |-- semester: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- _id: string (nullable = true)
 |-- code: string (nullable = true)
 |-- credits: integer (nullable = true)
 |-- description: string (nullable = true)
 |-- elective_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- key_assignments: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- course_name: string (nullable = true)
 |-- prerequisites: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- required_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
```

```
+----+------+-------+----------+--------+----+-------------+----+-----------+---
-----+----+------+------+-------+--------------------+--------------------+-----
-----------+--------------------+-------------+--------------------+
|term|course|section|enrollment|capacity| _id|academic_year|code|
name|semester|year|   _id|  code|credits|
description|elective_in_programs| key_assignments|
course_name|prerequisites|required_in_programs|
+----+------+-------+----------+--------+----+-------------+----+-----------+---
-----+----+------+------+-------+--------------------+--------------------+-----
-----------+--------------------+-------------+--------------------+
|1221|IST615|   M001|        22|      28|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST615|IST615|      3|Cloud services cr…|
[]|[project, paper]|    Cloud Management|           []|            [IS, DS]|
|1222|IST615|   M001|        19|      24|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST615|IST615|      3|Cloud services cr…|
[]|[project, paper]|    Cloud Management|           []|            [IS, DS]|
|1231|IST615|   M001|        21|      24|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST615|IST615|      3|Cloud services cr…|
[]|[project, paper]|    Cloud Management|           []|            [IS, DS]|
|1232|IST615|   M002|        20|      24|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST615|IST615|      3|Cloud services cr…|
[]|[project, paper]|    Cloud Management|           []|            [IS, DS]|
|1232|IST615|   M001|        21|      28|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST615|IST615|      3|Cloud services cr…|
[]|[project, paper]|    Cloud Management|           []|            [IS, DS]|
|1221|IST659|   M002|        20|      20|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1221|IST659|   M001|        20|      20|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1222|IST659|   M001|        24|      24|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1231|IST659|   M002|        20|      20|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1231|IST659|   M001|        20|      20|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1232|IST659|   M001|        20|      20|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST659|IST659|      3|Definition, devel…|                  []|
[project]|Data Administrati…|           []|            [IS, DS]|
|1221|IST687|   M002|        21|      24|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1221|IST687|   M001|        20|      20|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1222|IST687|   M002|        20|      20|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1222|IST687|   M001|        18|      20|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1231|IST687|   M002|        20|      24|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1231|IST687|   M001|        17|      20|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1232|IST687|   M001|        19|      24|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST687|IST687|      3|Introduces inform…|                [IS]|
[project, exam]|Introduction to D…|           []|                [DS]|
|1221|IST707|   M001|        28|      28|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST707|IST707|      3|General overview …|                [IS]|
[exam]|Applied Machine L…|     [IST687]|                [DS]|
|1231|IST707|   M001|        24|      24|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST707|IST707|      3|General overview …|                [IS]|
[exam]|Applied Machine L…|     [IST687]|                [DS]|
|1222|IST769|   M001|        19|      24|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST769|IST769|      3|Analyze relationa…|                [DS]|
[project, exam]|Advanced Big Data…|     [IST659]|                  []|
|1232|IST769|   M001|        20|      24|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST769|IST769|      3|Analyze relationa…|                [DS]|
[project, exam]|Advanced Big Data…|     [IST659]|                  []|
|1221|IST722|   M001|        25|      28|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST722|IST722|      3|Introduction to c…|                [IS]|
[project, exam]|    Data Warehousing|     [IST659]|                  []|
|1231|IST722|   M001|        23|      28|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST722|IST722|      3|Introduction to c…|                [IS]|
[project, exam]|    Data Warehousing|     [IST659]|                  []|
|1221|IST621|   M001|        22|      24|1221|    2021-2022|1221|  Fall 2021|
Fall|2021|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1222|IST621|   M002|        22|      24|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1222|IST621|   M001|        28|      28|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1231|IST621|   M001|        28|      28|1231|    2022-2023|1231|  Fall 2022|
Fall|2022|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1232|IST621|   M002|        21|      24|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1232|IST621|   M001|        28|      28|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST621|IST621|      3|Information and t…|                  []|
[paper]|Information Manag…|           []|                [IS]|
|1222|IST718|   M001|        28|      28|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST718|IST718|      3|A broad introduct…|                  []|
[project]|  Big Data Analytics|     [IST687]|                 [DS]|
|1232|IST718|   M001|         28|       28|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST718|IST718|      3|A broad introduct…|                  []|
[project]|  Big Data Analytics|     [IST687]|                 [DS]|
|1222|IST714|   M001|         17|       20|1222|    2021-2022|1222|Spring 2022|
Spring|2022|IST714|IST714|      3|Advanced, lab-bas…|              [IS, DS]|
[project]|  Cloud Architecture|     [IST615]|                   []|
|1232|IST714|   M001|         20|       24|1232|    2022-2023|1232|Spring 2023|
Spring|2023|IST714|IST714|      3|Advanced, lab-bas…|              [IS, DS]|
[project]|  Cloud Architecture|     [IST615]|                   []|
+----+------+-------+---------+--------+----+------------+----+-----------+---
-----+----+------+------+-------+--------------------+--------------------+-----
-----------+--------------------+-------------+--------------------+
```

1.0.5 Question 5

Use the `cassandra-driver` example from class to write python code to connect to cassandra from within Jupyter and create a keyspace named `ischooldb`. Design a cassandra table called `sections` to store the data from question 4. Appropriate key design is important! Please explain your justification for key below your table definition. Provide clear evidence that your table was created by querying the empty table in spark and use `printSchema ()` to show the schema.

[13]:
```
#5 create cassandra table for wide_sections

!pip install -q cassandra-driver

from cassandra.cluster import Cluster
with Cluster([cassandra_host]) as cluster:
    session = cluster.connect()
    session.execute("CREATE KEYSPACE IF NOT EXISTS ischooldb WITH replication={␣
 ↪'class': 'SimpleStrategy', 'replication_factor' : 1 };")
    table = '''
    CREATE TABLE IF NOT EXISTS ischooldb.sections (
        term INT,
        course TEXT,
        section TEXT,
        enrollment INT,
        capacity INT,
        academic_year TEXT,
        name TEXT,
        semester TEXT,
        year INT,
        credits INT,
        description TEXT,
        course_name TEXT,
        elective_in_programs  LIST<text> ,
        key_assignments  LIST<text>,
        prerequisites  LIST<text>,
        required_in_programs  LIST<text>,
        PRIMARY KEY((term, course),section));

    '''
    session.execute(table)
```

1.0.6 Question 6

Load the data frame you created in question 4 into the `cassandra` table you created in question 5. Demonstrate the data is in the table by querying back it with PySpark. Make sure you can run the code multiple times and each time it replaces the existing data.

[14]:

```
#6 load wide_sections into cassandra

section.write.format("org.apache.spark.sql.cassandra")\
  .mode("Append")\
  .option("table", "sections")\
  .option("keyspace", "ischooldb")\
  .save()
```

1.0.7 Question 7

Since we did not learn how to create a custom elasticsearch mapping, before you can load the data into `elasticsearch` you will need to flatten the nested data. For example, `course_is_elective_in_programs` should generate 2 columns `course_is_elective_for_IS` and `course_is_elective_for_DS`. You'll need to repeat this step for `course_is_required_in_programs`. Omit the `course_prerequisites` and `course_key_assignments` column.

[15]:

```
#7 flatten `course_is_elective_in_programs` and␣
 ↪`course_is_required_in_programs`
from pyspark.sql.functions import when, array_contains

df = section.
 ↪withColumn("course_is_elective_for_IS",when(array_contains(section["elective_in_programs"],
 ↪"IS"), "yes").otherwise("no"))

df = df.
 ↪withColumn("course_is_elective_for_DS",when(array_contains(df["elective_in_programs"],␣
 ↪"DS"), "yes").otherwise("no"))

df = df.
 ↪withColumn("course_is_required_for_IS",when(array_contains(df["required_in_programs"],␣
 ↪"IS"), "yes").otherwise("no"))
```

```
df = df.
 ↪withColumn("course_is_required_for_DS",when(array_contains(df["required_in_programs"],␣
 ↪"DS"), "yes").otherwise("no"))

df = df.drop("prerequisites", "key_assignments")

# Show the transformed DataFrame
df.show()
```

```
[Stage 73:=====================================================> (97 + 1) / 100]
+----+------+-------+----------+--------+-------------+-----------+--------+----
+-------+--------------------+--------------------+--------------------+--------
------------+-------------------------+-------------------------+---------------
----------+-------------------------+
|term|course|section|enrollment|capacity|academic_year|
name|semester|year|credits|         description|elective_in_programs|         co
urse_name|required_in_programs|course_is_elective_for_IS|course_is_elective_for_
DS|course_is_required_for_IS|course_is_required_for_DS|
+----+------+-------+----------+--------+-------------+-----------+--------+----
+-------+--------------------+--------------------+--------------------+--------
------------+-------------------------+-------------------------+---------------
----------+-------------------------+
|1221|IST615|   M001|        22|      28|    2021-2022|  Fall 2021|
Fall|2021|      3|Cloud services cr…|                  []|    Cloud
Management|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1222|IST615|   M001|        19|      24|    2021-2022|Spring 2022|
Spring|2022|      3|Cloud services cr…|                  []|    Cloud
Management|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1231|IST615|   M001|        21|      24|    2022-2023|  Fall 2022|
Fall|2022|      3|Cloud services cr…|                  []|    Cloud
Management|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1232|IST615|   M002|        20|      24|    2022-2023|Spring 2023|
Spring|2023|      3|Cloud services cr…|                  []|    Cloud
Management|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1232|IST615|   M001|        21|      28|    2022-2023|Spring 2023|
Spring|2023|      3|Cloud services cr…|                  []|    Cloud
Management|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1221|IST659|   M002|        20|      20|    2021-2022|  Fall 2021|
Fall|2021|      3|Definition, devel…|                  []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1221|IST659|   M001|        20|      20|    2021-2022|  Fall 2021|
Fall|2021|      3|Definition, devel…|                   []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1222|IST659|   M001|        24|      24|    2021-2022|Spring 2022|
Spring|2022|      3|Definition, devel…|                   []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1231|IST659|   M002|        20|      20|    2022-2023|  Fall 2022|
Fall|2022|      3|Definition, devel…|                   []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1231|IST659|   M001|        20|      20|    2022-2023|  Fall 2022|
Fall|2022|      3|Definition, devel…|                   []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1232|IST659|   M001|        20|      20|    2022-2023|Spring 2023|
Spring|2023|      3|Definition, devel…|                   []|Data
Administrati…|            [IS, DS]|                       no|
no|                      yes|                      yes|
|1221|IST687|   M002|        21|      24|    2021-2022|  Fall 2021|
Fall|2021|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1221|IST687|   M001|        20|      20|    2021-2022|  Fall 2021|
Fall|2021|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1222|IST687|   M002|        20|      20|    2021-2022|Spring 2022|
Spring|2022|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1222|IST687|   M001|        18|      20|    2021-2022|Spring 2022|
Spring|2022|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1231|IST687|   M002|        20|      24|    2022-2023|  Fall 2022|
Fall|2022|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1231|IST687|   M001|        17|      20|    2022-2023|  Fall 2022|
Fall|2022|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1232|IST687|   M001|        19|      24|    2022-2023|Spring 2023|
Spring|2023|      3|Introduces inform…|                 [IS]|Introduction to
D…|                [DS]|                      yes|                        no|
no|                      yes|
|1221|IST707|   M001|        28|      28|    2021-2022|  Fall 2021|
Fall|2021|      3|General overview …|                [IS]|Applied Machine
L…|                [DS]|                      yes|                       no|
no|                      yes|
|1231|IST707|   M001|        24|      24|    2022-2023|  Fall 2022|
Fall|2022|      3|General overview …|                [IS]|Applied Machine
L…|                [DS]|                      yes|                       no|
no|                      yes|
+----+------+-------+----------+--------+-------------+-----------+--------+----
+-------+--------------------+--------------------+--------------------+--------
------------+-------------------------+-------------------------+---------------
----------+-------------------------+
only showing top 20 rows
```

1.0.8 Question 8

Load the data frame you created in question 7 into `elasticsearch`, under the index `sections`. Demonstrate the data is in the index by querying back it with PySpark.

[16]:
```
#8 load wide_sections_flattened into elasticsearch
df.printSchema()
df.write.mode("Overwrite").format("es").save("sections/_doc")
```

```
root
 |-- term: integer (nullable = true)
 |-- course: string (nullable = true)
 |-- section: string (nullable = true)
 |-- enrollment: integer (nullable = true)
 |-- capacity: integer (nullable = true)
 |-- academic_year: string (nullable = true)
 |-- name: string (nullable = true)
 |-- semester: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- credits: integer (nullable = true)
 |-- description: string (nullable = true)
 |-- elective_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- course_name: string (nullable = true)
 |-- required_in_programs: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- course_is_elective_for_IS: string (nullable = false)
 |-- course_is_elective_for_DS: string (nullable = false)
 |-- course_is_required_for_IS: string (nullable = false)
 |-- course_is_required_for_DS: string (nullable = false)
```

1.0.9 Question 9

Similar to question 4, prepare the enrollments for loading into cassandra and elasticsearch with Spark or Spark SQL. For this wide table we want to include the same reference data for sections but include the student attributes and the program data associated with the student.

[17]:
```
#9 create wide_enrollments
programs = programs.withColumnRenamed("name","program_name")
students = students.withColumnRenamed("name","student_name")
students = students.withColumnRenamed("_id","lower_name")

combined2 = students.join(programs,students["program"] ==␣
 ↪programs["code"],"inner")
jnenroll = combined2.join(enrollment,combined2["lower_name"] ==␣
 ↪enrollment["student_id"],"inner")
#to avoid ambigous error:
jnenroll = jnenroll.withColumnRenamed("term","enroll_term")
jnenroll = jnenroll.withColumnRenamed("course","enroll_course")
jnenroll = jnenroll.withColumnRenamed("section","enroll_section")
jnenroll = jnenroll.withColumnRenamed("credits","total_credits")
jnenroll.show()
```

```
+------------------+------------------+-------+---+-----+-------------+-----------
-----+------------+--------------------+-------+-----------+-----------------+--
-----------+--------------+-----------------+-----+------------+
|        lower_name|
student_name|program|_id|code|total_credits|elective_courses|program_name|
required_courses|
type|enroll_term|course_enrollment|enroll_course|enroll_section|
student_id|grade|grade_points|
+------------------+------------------+-------+---+-----+-------------+-----------
-----+------------+--------------------+-------+-----------+-----------------+--
-----------+--------------+-----------------+-----+------------+
|          abbykuss|          Abby Kuss|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|       1232|                9|
IST621|           M001|          abbykuss|   A-|       3.667|
|          abbykuss|          Abby Kuss|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|       1231|                6|
IST722|           M001|          abbykuss|   A-|       3.667|
|          abbykuss|          Abby Kuss|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|       1221|               12|
IST707|           M001|          abbykuss|   A-|       3.667|
|          abbykuss|          Abby Kuss|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|       1221|               11|
IST687|           M002|          abbykuss|    A|         4.0|
|         abbykuss|         Abby Kuss|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1221|                 5|
IST659|          M001|         abbykuss|    A|          4.0|
|     aidenknowone|     Aiden Knowone|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1221|                 4|
IST621|          M001|     aidenknowone|    A|          4.0|
|         alfrecso|         Al Frecso|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                25|
IST621|          M001|         alfrecso|   C+|        2.333|
|         alfrecso|         Al Frecso|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1222|                19|
IST615|          M001|         alfrecso|    A|          4.0|
|          alkohol|          Al Kohol|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1222|                15|
IST621|          M001|          alkohol|    A|          4.0|
|amberwavesofgrain|Amber Wavesofgrain|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                10|
IST621|          M001|amberwavesofgrain|    A|          4.0|
|amberwavesofgrain|Amber Wavesofgrain|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1231|                 9|
IST722|          M001|amberwavesofgrain|   A-|        3.667|
|amberwavesofgrain|Amber Wavesofgrain|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1221|                 4|
IST659|          M001|amberwavesofgrain|   A-|        3.667|
|    anitasandwich|    Anita Sandwich|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                20|
IST718|          M001|    anitasandwich|    A|          4.0|
|    anitasandwich|    Anita Sandwich|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                12|
IST621|          M001|    anitasandwich|   B-|        2.667|
|    anitasandwich|    Anita Sandwich|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1221|                 5|
IST687|          M002|    anitasandwich|   A-|        3.667|
|      anitashower|      Anita Shower|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1222|                19|
IST687|          M002|      anitashower|    A|          4.0|
|      anitashower|      Anita Shower|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1222|                12|
IST621|          M002|      anitashower|    A|          4.0|
|       aprilfirst|       April First|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                25|
IST718|          M001|       aprilfirst|   A-|        3.667|
|       aprilfirst|       April First|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1232|                13|
IST621|          M001|       aprilfirst|    C|          2.0|
|       aprilfirst|       April First|     DS| DS|  DS|            34|[IST769,
IST714]|Data Science|[IST659, IST615, …|Masters|        1222|                14|
IST687|          M002|       aprilfirst|    A|          4.0|
+-----------------+------------------+-------+---+----+-------------+-----------
-----+------------+--------------------+-------+-----------+-----------------+--
-----------+--------------+-----------------+-----+------------+
only showing top 20 rows
```

[18]:

```
enrollments = jnenroll.join(section,(jnenroll["enroll_term"] ==␣
↪section["term"]) & \
                            (jnenroll["enroll_course"] == section["course"]) & \
                            (jnenroll["enroll_section"] ==␣
↪section["section"]),"inner")

enrollments.show()
```

```
+-----------------+------------------+-------+---+----+-------------+-----------
---------+-------------------+--------------------+-------+-----------+---------
--------+-------------+--------------+-----------------+-----+------------+----+
------+-------+----------+--------+-------------+---------+--------+----+-------
+--------------------+--------------------+---------------+----------------+----
---------+--------------------+
|       lower_name|      student_name|program|_id|code|total_credits|
elective_courses|       program_name|    required_courses|
type|enroll_term|course_enrollment|enroll_course|enroll_section|       student_i
d|grade|grade_points|term|course|section|enrollment|capacity|academic_year|
name|semester|year|credits|
description|elective_in_programs|key_assignments|
course_name|prerequisites|required_in_programs|
+-----------------+------------------+-------+---+----+-------------+-----------
---------+-------------------+--------------------+-------+-----------+---------
--------+-------------+--------------+-----------------+-----+------------+----+
------+-------+----------+--------+-------------+---------+--------+----+-------
+--------------------+--------------------+---------------+----------------+----
---------+--------------------+
|         abbykuss|         Abby Kuss|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
6|       IST722|          M001|         abbykuss|   A-|       3.667|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|amberwavesofgrain|Amber Wavesofgrain|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
9|       IST722|          M001|amberwavesofgrain|   A-|       3.667|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|  blanchedalmonds|  Blanche Dalmonds|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
3|       IST722|          M001|  blanchedalmonds|    A|         4.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|      chaselounge|      Chase Lounge|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
22|       IST722|          M001|      chaselounge|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|   dustindewinned|   Dustin DeWinned|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
21|       IST722|          M001|   dustindewinned|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|           ianewe|           Ian Ewe|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
11|       IST722|          M001|           ianewe|   B+|
3.333|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|         joyfulle|         Joy Fulle|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
1|       IST722|          M001|         joyfulle|    A|         4.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|         kurttain|         Kurt Tain|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
17|       IST722|          M001|         kurttain|    C|
2.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|         maximumm|         Max Imumm|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
8|       IST722|          M001|         maximumm|    A|         4.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|      mistyshores|      Misty Shores|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
7|       IST722|          M001|      mistyshores|    A|         4.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|       sherrywyne|       Sherry Wyne|     DS| DS|  DS|           34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
19|       IST722|          M001|       sherrywyne|   B-|
2.667|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|     sherylmytoyz|     Sheryl Mytoyz|     DS| DS|  DS|          34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
23|       IST722|          M001|     sherylmytoyz|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|         theodoor|         Theo Door|     DS| DS|  DS|          34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
16|       IST722|          M001|         theodoor|   B+|
3.333|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|      tuckandroll|      Tuck Androll|     DS| DS|  DS|          34|
[IST769, IST714]|       Data Science|[IST659, IST615, …|Masters|       1231|
5|       IST722|          M001|      tuckandroll|    B|         3.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|     carriemeehom|     Carrie Meehom|     IS| IS|  IS|          36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
18|       IST722|          M001|     carriemeehom|   B+|
3.333|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|        euronfyre|        Euron Fyre|     IS| IS|  IS|          36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
2|       IST722|          M001|        euronfyre|    A|         4.0|1231|IST722|
M001|        23|      28|    2022-2023|Fall 2022|    Fall|2022|
3|Introduction to c…|                [IS]|[project, exam]|Data Warehousing|
[IST659]|                  []|
|        frankklee|        Frank Klee|     IS| IS|  IS|          36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
20|       IST722|          M001|        frankklee|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|        hazeleyes|        Hazel Eyes|     IS| IS|  IS|          36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
14|       IST722|          M001|        hazeleyes|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|     holdenstrong|     Holden Strong|     IS| IS|  IS|          36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
12|       IST722|          M001|     holdenstrong|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
|       hughjapple|       Hugh Japple|     IS| IS|  IS|           36|[IST722,
IST714, …|Information Systems|[IST659, IST615, …|Masters|       1231|
15|       IST722|          M001|       hughjapple|    A|
4.0|1231|IST722|   M001|        23|      28|    2022-2023|Fall 2022|
Fall|2022|      3|Introduction to c…|                [IS]|[project, exam]|Data
Warehousing|     [IST659]|                  []|
+-----------------+------------------+-------+---+----+-------------+-----------
---------+-------------------+--------------------+-------+-----------+---------
--------+-------------+--------------+-----------------+-----+------------+----+
------+-------+----------+--------+-------------+---------+--------+----+-------
+--------------------+--------------------+---------------+----------------+----
---------+--------------------+
only showing top 20 rows
```

[19]:
```
enrollments.count()
```

[19]:
```
743
```

1.0.10 Question 10

Load the data frame you created in question 8 into `elasticsearch`, under the index `enrollments`. This time, just Omit all array types to make the problem simpler (`elective_courses`, `key_assignments`, `course_prerequisites`, etc…)

[20]:
```
#10 wide_enrollments to elastic search
enrollments=enrollments.drop("elective_courses",␣
 ↪"required_courses","elective_in_programs","key_assignments","prerequisites","required_in_pr
enrollments = enrollments.select ('lower_name',
 'student_name',
 'program',
 'total_credits',
 'program_name',
 'type',
 'enroll_term',
 'course_enrollment',
 'enroll_course',
 'enroll_section',
 'student_id',
 'grade',
 'grade_points',
 'term',
 'course',
 'section',
 'enrollment',
 'capacity',
 'academic_year',
 'name',
 'semester',
 'year',
 'credits',
 'description',
 'course_name')
```

[21]:
```
enrollments.write.mode("Overwrite").format("es").save("enrollments/_doc")
```

[22]:
```
#Cassandra
```

[23]:
```
!pip install -q cassandra-driver

from cassandra.cluster import Cluster
with Cluster([cassandra_host]) as cluster:
    session = cluster.connect()
    session.execute("CREATE KEYSPACE IF NOT EXISTS ischooldb WITH replication={␣
 ↪'class': 'SimpleStrategy', 'replication_factor' : 1 };")
    table = '''
    CREATE TABLE IF NOT EXISTS ischooldb.enrollments (
    lower_name TEXT,
    student_name TEXT,
    program TEXT,
    total_credits INT,
    program_name TEXT,
    type TEXT,
    enroll_term INT,
    course_enrollment INT,
    enroll_course TEXT,
    enroll_section TEXT,
    student_id TEXT,
    grade TEXT,
    grade_points DOUBLE,
    term INT,
    course TEXT,
    section TEXT,
    enrollment INT,
    capacity INT,
    academic_year TEXT,
    name TEXT,
    semester TEXT,
    year INT,
    credits INT,
    description TEXT,
    course_name TEXT,
    PRIMARY KEY((lower_name, term),section));
    '''
    session.execute(table)
```

[24]:

```
enrollments.write.format("org.apache.spark.sql.cassandra")\
  .mode("Append")\
  .option("table", "enrollments")\
  .option("keyspace", "ischooldb")\
  .save()
```

1.0.11 Question 11

Write spark to clear the `neo4j` database of all nodes and relationships.

[25]:

```
#11 reset neo4j database

cipher_ql = '''
MATCH (n)
DETACH DELETE n
'''
df = spark.createDataFrame(data = [{'row':1}])
df.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql) \
  .save()
```

1.0.12 Question 12

Load the `courses` and `program` data into `neo4j` as nodes. Exclude the `requirements`, `electives` and `prerequisites` from the node attributes. Demonstrate the data in `neo4j` by querying back it using one or more Cypher queries. NOTE: the Neo4J `name` attribute is what will display on the node bubbles.

[26]:

```
#12a load courses into Neo4j
print("Courses...")

cipher_ql = "Merge (c:Courses {code: event.code ,credits : event.credits␣
 ↪,description : event.description ,\
key_assignments: event.key_assignments , course_name : event.course_name})"

x = courses.
 ↪select("code","credits","description","key_assignments","course_name").
 ↪distinct()
x.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql) \
  .save()
```

```
Courses…
```

[27]:

```
query = '''
MATCH (c:Courses)
RETURN c.code AS code, c.credits AS credits, c.description AS description, c.
 ↪key_assignments AS key_assignments, c.course_name as course_name
'''
q12a = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query) \
                          .load()
q12a.show()
```

```
+------+-------+--------------------+----------------+--------------------+
|  code|credits|         description| key_assignments|         course_name|
+------+-------+--------------------+----------------+--------------------+
|IST718|      3|A broad introduct…|       [project]|  Big Data Analytics|
|IST722|      3|Introduction to c…| [project, exam]|    Data Warehousing|
|IST659|      3|Definition, devel…|       [project]|Data Administrati…|
|IST769|      3|Analyze relationa…| [project, exam]|Advanced Big Data…|
|IST621|      3|Information and t…|         [paper]|Information Manag…|
|IST707|      3|General overview …|          [exam]|Applied Machine L…|
|IST714|      3|Advanced, lab-bas…|       [project]|  Cloud Architecture|
|IST615|      3|Cloud services cr…|[project, paper]|    Cloud Management|
|IST687|      3|Introduces inform…| [project, exam]|Introduction to D…|
+------+-------+--------------------+----------------+--------------------+
```

[28]:

```
#12b load programs into neo4j
print("Programs...")
cipher_ql_p = '''
MERGE (p:Programs {code: event.code, credits: event.credits,program_name: event.
 ↪program_name, type: event.type})
'''


y = programs.select("code","credits","program_name","type").distinct()
```

```
y.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql_p) \
  .save()
```

Programs…

[29]:

```
query = '''
MATCH (p:Programs)
RETURN p.code AS code, p.credits AS credits, p.program_name AS program_name, p.
 ↪type AS type
'''
q12b = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query) \
                          .load()
q12b.show()
```

```
+----+-------+--------------------+-----------+
|code|credits|        program_name|       type|
+----+-------+--------------------+-----------+
|  DS|     34|        Data Science|    Masters|
| BDC|      9|Data Engineering …|Certificate|
| MLC|      9|Machine Learning …|Certificate|
| CCC|      9|Cloud Computing C…|Certificate|
|  IS|     36| Information Systems|    Masters|
+----+-------+--------------------+-----------+
```

[30]:

```
programs.columns
```

[30]:

```
['_id',
 'code',
 'credits',
 'elective_courses',
 'program_name',
 'required_courses',
 'type']
```

1.0.13 Question 13

Load the `requirements` and `electives` data into `neo4j` as relationships to the nodes you created in Question 12. Use the `program` data to form the `required` and `elective` course relationships. Demonstrate the relationships in `neo4j` are present by querying back it using one or more Cypher queries.

[31]:

```
#13a program course requirements

cipher_ql = """
MATCH (p:Programs {code: event.code})
UNWIND event.required_courses AS required_courses
MATCH (c:Courses {code: required_courses})
MERGE (p)-[:Requires {course: required_courses} ]->(c)
"""

programs.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql) \
  .save()
```

[32]:

```
query = '''
MATCH (p:Programs)-[r:Requires]->(c:Courses)
RETURN p.code AS ProgramCode, collect(c.code) AS RequiredCourses
'''
q13a = spark.read.format("org.neo4j.spark.DataSource") \
                           .option("url", bolt_url) \
                           .option("query",query) \
                           .load()
q13a.show()
```

```
+-----------+--------------------+
|ProgramCode|     RequiredCourses|
+-----------+--------------------+
|         DS|[IST718, IST659, …|
|        MLC|[IST718, IST707, …|
|        BDC|[IST722, IST659, …|
|         IS|[IST659, IST621, …|
|        CCC|[IST621, IST714, …|
+-----------+--------------------+
```

[33]:

```
#13b program course electives

cipher_ql = """
MATCH (p:Programs {code: event.code})
UNWIND event.elective_courses AS elective_courses
MATCH (c:Courses {code: elective_courses})
MERGE (p)-[:elective {course: elective_courses} ]->(c)
"""

programs.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql) \
  .save()
```

[34]:

```
query = '''
MATCH (p:Programs)-[e:elective]->(c:Courses)
RETURN p.code AS ProgramCode, collect(c.code) AS ElectiveCourses
'''
q13b = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query) \
                          .load()
q13b.show()
```

```
+-----------+--------------------+
|ProgramCode|     ElectiveCourses|
+-----------+--------------------+
|         IS|[IST722, IST714, …|
|         DS|    [IST769, IST714]|
+-----------+--------------------+
```

1.0.14 Question 14

Load the `prerequisites` into `neo4j` as relationships to the `course` nodes you created in Question 12. Demonstrate the relationships in `neo4j` are present by querying back it using one or more Cypher queries.

[35]:

```
cipher_ql = """
MATCH (c:Courses {code: event.code})
UNWIND event.prerequisites AS prerequisites
MATCH (n:Courses {code: prerequisites})
MERGE (c)-[:prerequisites {course: prerequisites} ]->(n)
"""

courses.write.format("org.neo4j.spark.DataSource").mode("Overwrite") \
  .option("url", bolt_url) \
  .option("query",cipher_ql) \
  .save()
```

[36]:

```
query = '''
MATCH (c:Courses)-[r:prerequisites]->(n:Courses)
RETURN c.code AS CourseCode, collect(n.code) AS Prerequisites
'''
q14 = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query) \
                          .load()
q14.show()
```

```
+----------+-------------+
|CourseCode|Prerequisites|
+----------+-------------+
|    IST722|     [IST659]|
|    IST769|     [IST659]|
|    IST714|     [IST615]|
|    IST707|     [IST687]|
|    IST718|     [IST687]|
+----------+-------------+
```

1.0.15 Question 15

Write a Cypher query to display courses which are required by both the `IS` and `DS` programs.

[37]:

```
#15 Cypher query courses required in DS and IS

query = '''
MATCH (is:Programs {code: "IS"})-[:Requires]->(c:Courses)<-[:Requires]-(ds:
 ↪Programs {code: "DS"})
RETURN c.code as course_id,c.course_name as course_name,c.credits as␣
 ↪course_credits,c.description as course_description
'''
q15 = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query) \
                          .load()
q15.show()
```

```
+---------+--------------------+--------------+--------------------+
|course_id|         course_name|course_credits|  course_description|
+---------+--------------------+--------------+--------------------+
|   IST615|    Cloud Management|             3|Cloud services cr…|
|   IST659|Data Administrati…|             3|Definition, devel…|
+---------+--------------------+--------------+--------------------+
```

1.0.16 Question 16

Write a Cypher query to retrieve the `course code`, `course title`, and the count of programs the course is a requirement in. Write as a Cypher query but retrieve the output as a Spark Dataframe.

[38]:

```
#16 Cypher to spark table

query_t = '''
MATCH (p:Programs)-[:Requires]->(c:Courses)
RETURN c.code AS CourseCode, c.course_name AS CourseTitle, count(p) AS␣
 ↪ProgramCount
```

```
'''
q16 = spark.read.format("org.neo4j.spark.DataSource") \
                          .option("url", bolt_url) \
                          .option("query",query_t) \
                          .load()
q16.show()
```

```
+----------+--------------------+------------+
|CourseCode|         CourseTitle|ProgramCount|
+----------+--------------------+------------+
|    IST718|  Big Data Analytics|           2|
|    IST722|    Data Warehousing|           1|
|    IST659|Data Administrati…|           3|
|    IST769|Advanced Big Data…|           1|
|    IST621|Information Manag…|           2|
|    IST707|Applied Machine L…|           2|
|    IST714|  Cloud Architecture|           1|
|    IST615|    Cloud Management|           3|
|    IST687|Introduction to D…|           2|
+----------+--------------------+------------+
```

1.0.17 Questions 17,18,19 and 20

These are not spark questions as they use kibana.