Spark Assingment: CSVToDFAndDS

Michael Edelman

May 2022

1 Phase 1

1.1 df.show the first 5 books from the DataFrame ingestion: i.e., Dataset < Row >

```
+---+--------+--------------------+-----------+--------------------+
| id|authorId|               title|releaseDate|                link|
+---+--------+--------------------+-----------+--------------------+
|  1|       1|Fantastic Beasts ...|   11/18/16|http://amzn.to/2k...|
|  2|       1|Harry Potter and ...|    10/6/15|http://amzn.to/2l...|
|  3|       1|The Tales of Beed...|    12/4/08|http://amzn.to/2k...|
|  4|       1|Harry Potter and ...|    10/4/16|http://amzn.to/2k...|
|  5|       2|Informix 12.10 on...|    4/23/17|http://amzn.to/2i...|
+---+--------+--------------------+-----------+--------------------+
ıly showing top 5 rows
```

1.2 df.printSchema that DataFrame

```
root
 |-- id: string (nullable = true)
 |-- authorId: string (nullable = true)
 |-- title: string (nullable = true)
 |-- releaseDate: string (nullable = true)
 |-- link: string (nullable = true)
```

1.3 df.show all rows in that DataFrame "where the authorId equals 1". This will demonstrate the generic-but-powerful "DataFrame" API.

```
+---+--------+--------------------+-----------+--------------------+
| id|authorId|               title|releaseDate|                link|
+---+--------+--------------------+-----------+--------------------+
|  1|       1|Fantastic Beasts ...|   11/18/16|http://amzn.to/2k...|
|  2|       1|Harry Potter and ...|    10/6/15|http://amzn.to/2l...|
|  3|       1|The Tales of Beed...|    12/4/08|http://amzn.to/2k...|
|  4|       1|Harry Potter and ...|    10/4/16|http://amzn.to/2k...|
+---+--------+--------------------+-----------+--------------------+
```

2 Phase 2

2.1 df.show(5,17) the books from the DataFrame ingestion: i.e., Dataset < Book >

```
+--------+---+-----------------+-----------------+-----------------+
|authorId| id|             link|      releaseDate|            title|
+--------+---+-----------------+-----------------+-----------------+
|       1|  1|http://amzn.to...|[18, 0, 0, 11,...|Fantastic Beas...|
|       1|  2|http://amzn.to...|[6, 0, 0, 10, ...|Harry Potter a...|
|       1|  3|http://amzn.to...|[4, 0, 0, 0, 0...|The Tales of B...|
|       1|  4|http://amzn.to...|[4, 0, 0, 10, ...|Harry Potter a...|
|       2|  5|http://amzn.to...|[23, 0, 0, 4, ...|Informix 12.10...|
+--------+---+-----------------+-----------------+-----------------+
only showing top 5 rows
```

2.2 df.printSchema that Dataset

```
root
 |-- authorId: integer (nullable = true)
 |-- id: integer (nullable = true)
 |-- link: string (nullable = true)
 |-- releaseDate: struct (nullable = true)
 |    |-- date: integer (nullable = true)
 |    |-- hours: integer (nullable = true)
 |    |-- minutes: integer (nullable = true)
 |    |-- month: integer (nullable = true)
 |    |-- seconds: integer (nullable = true)
 |    |-- time: long (nullable = true)
 |    |-- year: integer (nullable = true)
 |-- title: string (nullable = true)
```

2.3 Iterate over all the Book instances of this Dataset, doing a println of that instance. This will demonstrate the "strong typing" of the "Dataset" API.

```
Book: id=1,authorId=1, title=Fantastic Beasts and Where to Find Them: The Original Screenplay
Book: id=2,authorId=1, title=Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter, Book 1
Book: id=3,authorId=1, title=The Tales of Beedle the Bard, Standard Edition (Harry Potter)
Book: id=4,authorId=1, title=Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter, Book
Book: id=5,authorId=2, title=Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee
Book: id=6,authorId=2, title=Development Tools in 2006: any Room for a 4GL-style Language?: An independent study
Book: id=7,authorId=3, title=Adventures of Huckleberry Finn
Book: id=8,authorId=3, title=A Connecticut Yankee in King Arthur's Court
Book: id=10,authorId=4, title=Jacques le Fataliste
Book: id=11,authorId=4, title=Diderot Encyclopedia: The Complete Illustrations 1762-1777
Book: id=12,authorId=0, title=A Woman in Berlin
Book: id=13,authorId=6, title=Spring Boot in Action
Book: id=14,authorId=6, title=Spring in Action: Covers Spring 4
Book: id=15,authorId=7, title=Soft Skills: The software developer's life manual
Book: id=16,authorId=8, title=Of Mice and Men
Book: id=17,authorId=9, title=Java
Book: id=18,authorId=12, title=Hamlet
Book: id=19,authorId=13, title=Penses
Book: id=20,authorId=14, title=Fables choisies, mises en vers par M. de La Fontaine
Book: id=21,authorId=15, title=Discourse on Method and Meditations on First Philosophy
Book: id=22,authorId=12, title=Twelfth Night
Book: id=23,authorId=12, title=Macbeths
```

2.4 show all rows in that DataFrame "where the authorId equals 1

```
+--------+---+--------------------+--------------------+--------------------+
|authorId| id|                link|         releaseDate|               title|
+--------+---+--------------------+--------------------+--------------------+
|       1|  1|http://amzn.to/2k...|[18, 0, 0, 11, 0,...|Fantastic Beasts ...|
|       1|  2|http://amzn.to/2l...|[6, 0, 0, 10, 0, ...|Harry Potter and ...|
|       1|  3|http://amzn.to/2k...|[4, 0, 0, 0, 0, 4...|The Tales of Beed...|
|       1|  4|http://amzn.to/2k...|[4, 0, 0, 10, 0, ...|Harry Potter and ...|
+--------+---+--------------------+--------------------+--------------------+
```

3 Phase 3

3.1 As you'll see for yourself, the releaseDate representation from the previous phase looks bizarre in "schema" form. This is because Spark's mapping of a Java Data is a set of nested fields. You must convert this state into a single, new column for this DataFrame called releaseDateAsString. The type of this new column is String, and the format of values for this column is YYYY-MM-DD.

authorId	id	link	releaseDate	title	releaseDateAsString
1	1	http://amzn.to/2k...	[18, 0, 0, 11, 0,...	Fantastic Beasts ...	2016-11-18
1	2	http://amzn.to/2l...	[6, 0, 0, 10, 0, ...	Harry Potter and ...	2015-10-6
1	3	http://amzn.to/2k...	[4, 0, 0, 0, 0, 4...	The Tales of Beed...	209-0-4
1	4	http://amzn.to/2k...	[4, 0, 0, 10, 0, ...	Harry Potter and ...	2016-10-4
2	5	http://amzn.to/2i...	[23, 0, 0, 4, 0, ...	Informix 12.10 on...	2017-4-23
2	6	http://amzn.to/2v...	[28, 0, 0, 0, 0, ...	Development Tools...	2017-0-28
3	7	http://amzn.to/2w...	[26, 0, 0, 5, 0, ...	Adventures of Huc...	1994-5-26
3	8	http://amzn.to/2x...	[17, 0, 0, 6, 0, ...	A Connecticut Yan...	2017-6-17
4	10	http://amzn.to/2u...	[1, 0, 0, 3, 0, 4...	Jacques le Fataliste	200-3-1
4	11	http://amzn.to/2i...	null	Diderot Encyclope...	
0	12	http://amzn.to/2i...	[11, 0, 0, 7, 0, ...	A Woman in Berlin	206-7-11
6	13	http://amzn.to/2h...	[3, 0, 0, 1, 0, 6...	Spring Boot in Ac...	2016-1-3
6	14	http://amzn.to/2y...	[28, 0, 0, 11, 0,...	Spring in Action:...	2014-11-28
7	15	http://amzn.to/2z...	[29, 0, 0, 0, 0, ...	Soft Skills: The ...	2015-0-29
8	16	http://amzn.to/2z...	null	Of Mice and Men	
9	17	http://amzn.to/2i...	[28, 0, 0, 8, 0, ...	Java	2014-8-28
12	18	http://amzn.to/2y...	[8, 0, 0, 6, 0, 6...	Hamlet	2012-6-8
13	19	http://amzn.to/2j...	[9, 0, 0, 6, 0, 6...	Penses	2013-6-9
14	20	http://amzn.to/2y...	[1, 0, 0, 9, 0, 6...	Fables choisies, ...	1999-9-1
15	21	http://amzn.to/2h...	[15, 0, 0, 6, 0, ...	Discourse on Meth...	1999-6-15

only showing top 20 rows

3.2 Create a column releaseDateAsDate of type Date from releaseDateAsString, then df.drop the releaseDateAsString column.

authorId	id	link	releaseDate	title	releaseDateAsDate
1	1	http://amzn.to/2k...	[18, 0, 0, 11, 0,...	Fantastic Beasts ...	2016-11-18
1	2	http://amzn.to/2l...	[6, 0, 0, 10, 0, ...	Harry Potter and ...	2015-10-06
1	3	http://amzn.to/2k...	[4, 0, 0, 0, 0, 4...	The Tales of Beed...	null
1	4	http://amzn.to/2k...	[4, 0, 0, 10, 0, ...	Harry Potter and ...	2016-10-04
2	5	http://amzn.to/2i...	[23, 0, 0, 4, 0, ...	Informix 12.10 on...	2017-04-23
2	6	http://amzn.to/2v...	[28, 0, 0, 0, 0, ...	Development Tools...	null
3	7	http://amzn.to/2w...	[26, 0, 0, 5, 0, ...	Adventures of Huc...	1994-05-26
3	8	http://amzn.to/2x...	[17, 0, 0, 6, 0, ...	A Connecticut Yan...	2017-06-17
4	10	http://amzn.to/2u...	[1, 0, 0, 3, 0, 4...	Jacques le Fataliste	null
4	11	http://amzn.to/2i...	null	Diderot Encyclope...	null
0	12	http://amzn.to/2i...	[11, 0, 0, 7, 0, ...	A Woman in Berlin	null
6	13	http://amzn.to/2h...	[3, 0, 0, 1, 0, 6...	Spring Boot in Ac...	2016-01-03
6	14	http://amzn.to/2y...	[28, 0, 0, 11, 0,...	Spring in Action:...	2014-11-28
7	15	http://amzn.to/2z...	[29, 0, 0, 0, 0, ...	Soft Skills: The ...	null
8	16	http://amzn.to/2z...	null	Of Mice and Men	null
9	17	http://amzn.to/2i...	[28, 0, 0, 8, 0, ...	Java	2014-08-28
12	18	http://amzn.to/2y...	[8, 0, 0, 6, 0, 6...	Hamlet	2012-06-08
13	19	http://amzn.to/2j...	[9, 0, 0, 6, 0, 6...	Penses	2013-06-09
14	20	http://amzn.to/2y...	[1, 0, 0, 9, 0, 6...	Fables choisies, ...	1999-09-01
15	21	http://amzn.to/2h...	[15, 0, 0, 6, 0, ...	Discourse on Meth...	1999-06-15

only showing top 20 rows

3.3 df.drop the releaseDate column from the previous phases (it's no longer necessary, since you've created the releaseDateAsDate column).

3.4 df.show(5,17) the books from the DataFrame that's been transformed from the Dataset < Book >

```
+--------+---+----------------+----------------+----------------+
|authorId| id|            link|           title|releaseDateAsDate|
+--------+---+----------------+----------------+----------------+
|       1|  1|http://amzn.to...|Fantastic Beas...|      2016-11-18|
|       1|  2|http://amzn.to...|Harry Potter a...|      2015-10-06|
|       1|  3|http://amzn.to...|The Tales of B...|            null|
|       1|  4|http://amzn.to...|Harry Potter a...|      2016-10-04|
|       2|  5|http://amzn.to...|Informix 12.10...|      2017-04-23|
+--------+---+----------------+----------------+----------------+
only showing top 5 rows
```

3.5 df.printSchema that DataFrame: that will demonstrate how the DataFrame has considerably more information about the underlying schema than it had in Phase 1. The extra information comes from the "strong typing" representation performed in Phase 2.

```
root
 |-- authorId: integer (nullable = true)
 |-- id: integer (nullable = true)
 |-- link: string (nullable = true)
 |-- title: string (nullable = true)
 |-- releaseDateAsDate: date (nullable = true)
```

3.6 Sort this DataFrame by descending values of releaseDateAsDate, then display the contents using show(20, 15).

```
+--------+---+--------------+--------------+-----------------+
|authorId| id|          link|         title|releaseDateAsDate|
+--------+---+--------------+--------------+-----------------+
|       3|  8|http://amzn....|A Connecticu...|       2017-06-17|
|       2|  5|http://amzn....|Informix 12....|       2017-04-23|
|       1|  1|http://amzn....|Fantastic Be...|       2016-11-18|
|       1|  4|http://amzn....|Harry Potter...|       2016-10-04|
|       6| 13|http://amzn....|Spring Boot ...|       2016-01-03|
|       1|  2|http://amzn....|Harry Potter...|       2015-10-06|
|       6| 14|http://amzn....|Spring in Ac...|       2014-11-28|
|       9| 17|http://amzn....|          Java|       2014-08-28|
|      13| 19|http://amzn....|        Penses|       2013-06-09|
|      12| 18|http://amzn....|        Hamlet|       2012-06-08|
|      12| 23|http://amzn....|      Macbeths|       2003-07-01|
|      12| 22|http://amzn....|  Twelfth Night|       2003-07-01|
|      14| 20|http://amzn....|Fables chois...|       1999-09-01|
|      15| 21|http://amzn....|Discourse on...|       1999-06-15|
|       3|  7|http://amzn....|Adventures o...|       1994-05-26|
|       1|  3|http://amzn....|The Tales of...|             null|
|       2|  6|http://amzn....|Development ...|             null|
|       4| 10|http://amzn....|Jacques le F...|             null|
|       0| 12|http://amzn....|A Woman in B...|             null|
|       4| 11|http://amzn....|Diderot Ency...|             null|
+--------+---+--------------+--------------+-----------------+
only showing top 20 rows
```