

## Correção: Campo 'Date Added'

### 01. DADOS GERAIS

Linguagem de Programação:	Scala
Build:	SBT (Scala Build Tool)
Categoria:	Correção

### 02. ERRO/DEFEITO

1. O Valor da Chave 'Date Added' nos documentos gerados no mongoDB pelo projeto SparkCovid estão com todos os valores vazios, conforme imagem abaixo:

```
_id: ObjectId('647731eace08152f3a4b3031')
Abstract: "Crowdlending platforms have gained importance in recent years due to d..."
Accession Number: ""
AlternateId: ""
Authors: "José María, Ferrer, Klaus, Ulrich, González-Tejero Cristina, Blanco, M..."
CovNum: "2095573"
Database: "Science Direct"
Date Added: ""
DOI: "10.1016/j.jbusres.2022.113433"
FulltextLink: "https://doi.org/10.1016/j.jbusres.2022.113433"
Issue: ""
Journal: "Journal of Business Research"
Keywords: "Crowdlending|Crowdfunding|Fintech|Finance|Covid-19"
KJD: ""
Language: "en"
Pages: "113433"
PMID: ""
PublishDate: ""
Published Month: ""
Published Year: "2022"
SCIELO: ""
Tags: ""
Title: "Investors' confidence in the crowdlending platform and the impact of C..."
UNKNOWN: ""
Volume: ""
WOS: ""
_updd: 2023-05-31T03:00:00.000+00:00
_upddSrc: 2022-11-21T03:00:00.000+00:00
```

2. A incorreta definição do valor foi encontrada através do modulo 'ScalaMongoDiff' comparando duas coleções no mongoDB. Uma coleção incluída pelo SparkCovid e outra pelo MongoDBMigrations lendo diretamente as planilhas do PENTAH0.

### 03. ESPERADO

1. É esperado que o valor da chave 'Date Added' tenha o conteúdo preenchido, conforme a imagem abaixo:

```
_id: ObjectId('6477369791632340a4cc28b1')
WOS: "000890338200041"
FulltextLink: "https://doi.org/10.47989/irisic2235"
PMID: ""
SCIELO: ""
AlternateId: ""
Issue: ""
Authors: "Caidi, N.| Ekmekcioglu, C.| Jamali, R.| Chandra, P."
Journal: "Information Research-an International Electronic Journal"
KJD: ""
Published Month: ""
Title: "Socially-distant fasting: information practices of young muslims durin..."
CovNum: "2146753"
Volume: "27"
Pages: ""
Database: "Web of Science"
PublishDate: ""
Language: "en"
Date Added: "20/12/2022"
Published Year: "2022"
Abstract: "Introduction. The COVID-19 pandemic has forced people to reimagine how..."
Keywords: ""
Tags: ""
DOI: "10.47989/irisic2235"
UNKNOWN: ""
Accession Number: "WOS:000890338200041"
```

## 04. CORREÇÃO NO CÓDIGO

1. Criado o objeto Ddate e o metodo fix transformando o formato da data, conforme abaixo:

```
Ddate.scala x
1 package sp
2
3 import org.apache.spark.sql.expressions.UserDefinedFunction
4 import org.apache.spark.sql.functions.udf
5
6 new *
7 object Ddate {
8   new *
9   def fix(in: String): String = {
10
11     val parts: Array[String] = in.split("-")
12
13     val day = parts(0)
14     val month = parts(1)
15     val year = "20" + parts(2)
16
17     s"$day/$month/$year"
18   }
19
20   val _udf: UserDefinedFunction = udf((in: String) => fix(in))
21 }
```

2. Na variável df3 modificado para obter o resultado do campo D date e renomeado em seguida pra Date Added., conforme abaixo:

```
SparkCovidWho.scala x
46 .withColumn("Journal", Journal._udf(col("Journal")))
47 .withColumn("Keywords", Keywords._udf(col("Keywords")))
48 .withColumn("Language", Language._udf(col("Language")))
49 .withColumn("Pages", Pages._udf(col("Pages")))
50 .withColumn("Published Month", PublishedMonth._udf(col("Publishdate")))
51 .withColumn("Published Year", PublishedYear._udf(col("Year")))
52 .withColumn("Title", Title._udf(col("Title")))
53 .withColumn("Volume", Volume._udf(col("Volume")))
54 .withColumn("_upddSrc", UpddSrc._udf(col("_updd")))
55 .withColumn("_updd", lit(date))
56 .withColumn("PMID", PMID._udf(col("Accession number")))
57 .withColumn("WOS", WOS._udf(col("Accession number")))
58 .withColumn("KJD", KJD._udf(col("Accession number")))
59 .withColumn("SCIELO", SCIELO._udf(col("Accession number")))
60 .withColumn("UNKNOWN", UNKNOWN._udf(col("Accession number")))
61 df1.printSchema()
62
63 val df2: DataFrame =
64   df1.withColumn("FulltextLink", FulltextLink._udf(col("Doi"), col("PMID")))
65   df2.printSchema()
66
67 val df3: DataFrame = df2.select("Abstract", "Accession Number", "AlternateId", "Authors", "CovNum",
68   "Database", "D_date", "DOI", "FulltextLink", "Issue", "Journal", "Keywords", "KJD", "Language", "Pages", "PMID",
69   "PublishDate", "Published Month", "Published Year", "SCIELO", "Tags", "Title", "UNKNOWN", "Volume", "WOS", "_updd",
70   "_upddSrc").withColumnRenamed("D_date", "Date Added")
71 df3.printSchema()
72 df3.show(3, truncate = true)
```