DATA WRANGLING USING PYSPARK

Project Aim and Objectives

Aim

This project aims to type-cast and transform a log data using the "pyspark" module in Python.

Objectives

- To access the zipped log data from a specified external source
- To extract the zipped data using the "tarfile" module
- To read the data using the "pyspark" module
- To transform the raw data using the "pyspark" module

Data Description

The raw data, namely "BGL.log", contains lines of information of system control failure prompts.

```
- 1117838570 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.50.363779 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838570 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.50.527847 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838570 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.50.675872 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838570 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.50.823719 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838570 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.50.982731 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.131467 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.293532 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.428563 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.601412 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.749199 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838571 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.51.885834 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.041388 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.199063 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.345821 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.493353 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.638135 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.807927 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838572 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.52.951717 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.125780 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.276129 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.414979 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.573391 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.722611 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838573 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.53.870042 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.019446 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.161937 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.309788 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.532078 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.672029 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.822857 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838574 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.54.958702 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
- 1117838575 2005.06.03 R02-M1-N0-C:J12-U11 2005-06-03-15.42.55.114809 R02-M1-N0-C:J12-U11 RAS KERNEL INFO instruction cache parity error corrected
```

Figure 1: Snippet of raw data

The segment of data shown in *figure 1* contains the system failure prompts in the first line of which,

- "-" is the "Alert message flag" for a specific prompt

- “1117838570” is the “Timestamp” for a specific prompt
- “2005.06.03” is the “Date” on which a specific prompt is generated
- “R02-M1-N0-C:J12-U11” is the “Node” that is responsible for a particular prompt
- “2005-06-03 15:42:50.363779” is the date, in a typical date-time format, on which a specific prompt is thrown by a system
- “R02-M1-N0-C:J12-U11” is the confirmation of the “Node”, to be denoted as “Node (repeated), for which a system has encountered an error
- “RAS” is the “Message Type”
- “KERNEL” is the “System Component” that is associated with the error
- “INFO” is the “Level” of the error prompt
- “instruction cache parity error corrected” is the “Message Content”

Data are stored in the remaining lines of log in the same format. This data has to be transformed using the “pyspark” module so that it gets the following schema.

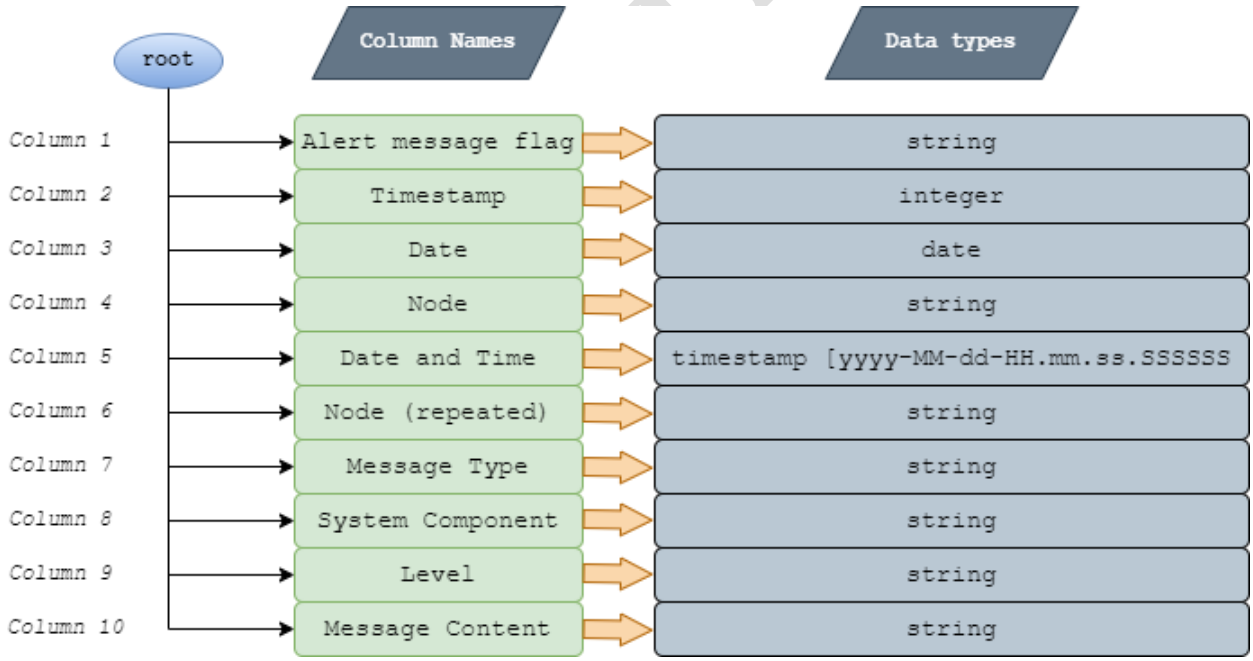

Figure 2: Dataframe schema

The length of each of the line varies and needs to be oriented in ten columns where each line should represent each row in the pyspark dataframe.

Problem Statement

This raw data could easily be restructured by separating each content (or word) of each line by the “space” character but this would also separate the “Message Contents” into more than one column like the following.

Incorrect Format

The “Message Content” column should have contained the entire prompt “8 floating point alignment exceptions” but instead it has “8” now as the raw data has been separated by a “space” character.

Alert message flag	Timestamp	Date	Node	Date and time	Node (repeated)	Message Type	System Component	Level	Message Content
-	1121902393	2005.07.20	R22-M1-N2-C:J13-U11	2005-07-20-16.33.13.751938	R22-M1-N2-C:J13-U11	RAS	KERNEL	INFO	8

Correct Format

This is the format the restructured data should have to which this data-preprocessing project is aimed at.

Alert message flag	Timestamp	Date	Node	Date and time	Node (repeated)	Message Type	System Component	Level	Message

									Content
-	1121902393	2005.07.20	R22-M1-N2-C:J13-U11	2005-07-20-16.33.13.751938	R22-M1-N2-C:J13-U11	RAS	KERNEL	INFO	8 floating point alignment exceptions

Other than restructuring, the variables (or the columns) of the dataframe also need to be type-casted as the entire raw data is of "string" type (object type).

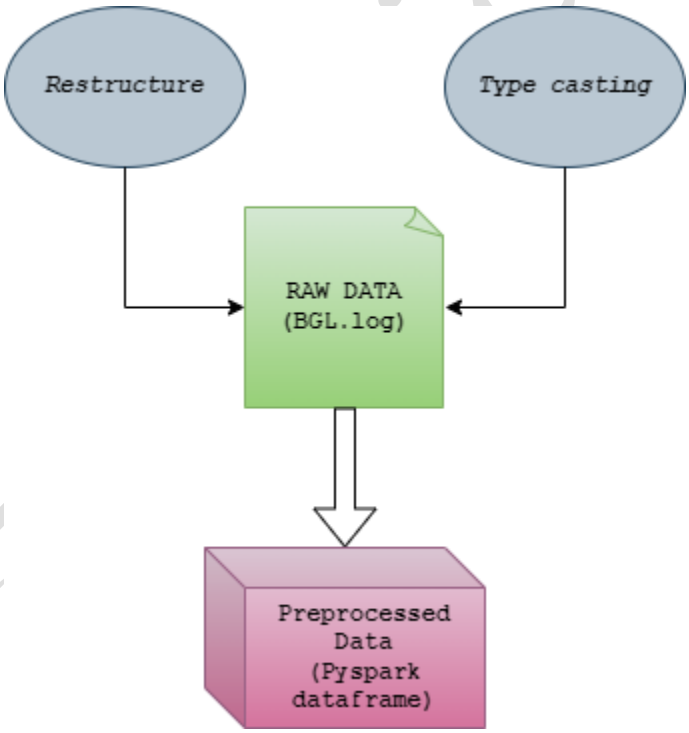

Figure 3: Diagrammatic representation of problem statement and solution

The solution to the aforementioned problem has been provided in the following section.

Solution

The “space” character has been used here as the delimiter to divide each line of the log data into ten substrings. From left to right the control flows and it defines each substring as the “word(s)” it encounters before a “space” character. Using the “getItem()” function the ten substrings are assigned to ten different columns.

```
dat = dat.withColumn("Message Content", split(col("value"), " ", 10).getItem(9))
```

Figure 4: Method employed for restructuring the data

Here,

Statement Component	Explanation
dat	Name of the variable that stores the dataframe
“.withColumn()”	Method used to add a new column to the dataframe “dat”
“Message Content”	Is the name of the new column
split(col("value"), " ", 10)	The “split()” function, used for separating each line, takes three inputs – i. The dataframe column, “value”, whose content needs to be separated ii. A delimiter, here a space character iii. The number of substrings in which the dataframe needs to be divided

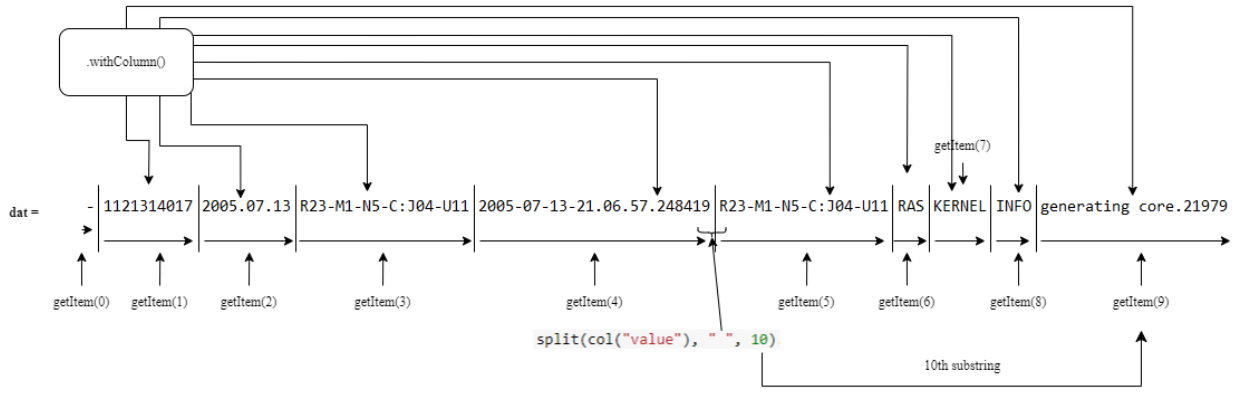

Figure 5: Code explanation

```
+------------------+----------+----------+------------------+-------------------------+------------------+------------+-----
----------+-----+---------------------------------------+
|Alert message flag|Timestamp |Date      |Node              |Date and Time            |Node (repeated)   |Message Type|Syste
m Component|Level|Message Content                        |
+------------------+----------+----------+------------------+-------------------------+------------------+------------+-----
----------+-----+---------------------------------------+
|-                 |1117838570|2005-06-03|R02-M1-N0-C:J12-U11|2005-06-03 15:42:50.363779|R02-M1-N0-C:J12-U11|RAS         |KERNE
L          |INFO |instruction cache parity error corrected|
|-                 |1117838570|2005-06-03|R02-M1-N0-C:J12-U11|2005-06-03 15:42:50.527847|R02-M1-N0-C:J12-U11|RAS         |KERNE
L          |INFO |instruction cache parity error corrected|
+------------------+----------+----------+------------------+-------------------------+------------------+------------+-----
----------+-----+---------------------------------------+
only showing top 2 rows
```

Figure 6: Transformed data

```
root
 |-- Alert message flag: string (nullable = true)
 |-- Timestamp: integer (nullable = true)
 |-- Date: date (nullable = true)
 |-- Node: string (nullable = true)
 |-- Date and Time: timestamp (nullable = true)
 |-- Node (repeated): string (nullable = true)
 |-- Message Type: string (nullable = true)
 |-- System Component: string (nullable = true)
 |-- Level: string (nullable = true)
 |-- Message Content: string (nullable = true)
```

Figure 7: Dataframe schema