Homework 2 - Distributed Data Managment

Install findspark, pyspark in case it is not installed - if running on Colab.

You can copy the whole notebook to your Google account and work on it on Colab via:

File -> Save a copy in Drive -> Open the copied notebook

In [3]:

```
# * When using the Docker workspace do not run this step *
IM_RUNNNING_ON_COLAB = False
if IM_RUNNNING_ON_COLAB:

  !pip install --force-reinstall pyspark==3.2
  !pip install findspark
```

Uplaod the data from Moodle, it's a zip file so simply unzip it

In [4]:

```
!unzip /content/random_data.parquet.zip
```

```
Archive:  /content/random_data.parquet.zip
   creating: random_data.parquet/
 extracting: random_data.parquet/.part-00000-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00001-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00002-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00003-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00004-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00005-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00006-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00007-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00008-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00009-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00010-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00011-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00012-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00013-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00014-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00015-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00016-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00017-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00018-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00019-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00020-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
```

```
appy.parquet.crc
 extracting: random_data.parquet/.part-00021-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00022-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00023-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00024-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00025-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00026-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00027-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00028-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00029-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00030-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00031-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00032-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00033-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00034-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00035-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00036-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00037-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00038-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00039-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00040-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00041-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00042-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00043-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00044-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00045-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00046-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00047-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00048-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00049-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00050-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00051-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00052-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00053-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00054-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00055-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00056-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
```

```
appy.parquet.crc
 extracting: random_data.parquet/.part-00057-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00058-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00059-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00060-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00061-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00062-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00063-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00064-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00065-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00066-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00067-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00068-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00069-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00070-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00071-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00072-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00073-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00074-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00075-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00076-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00077-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00078-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00079-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00080-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00081-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00082-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00083-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00084-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00085-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00086-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00087-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00088-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00089-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00090-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00091-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00092-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
```

```
appy.parquet.crc
 extracting: random_data.parquet/.part-00093-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00094-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00095-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00096-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00097-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00098-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00099-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00100-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00101-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00102-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00103-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00104-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00105-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00106-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00107-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00108-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00109-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00110-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00111-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00112-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00113-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00114-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00115-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00116-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00117-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00118-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00119-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00120-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00121-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00122-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00123-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00124-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00125-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00126-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00127-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00128-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
```

```
appy.parquet.crc
 extracting: random_data.parquet/.part-00129-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00130-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00131-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00132-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00133-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00134-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00135-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00136-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00137-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00138-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00139-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00140-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00141-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00142-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00143-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00144-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00145-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00146-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00147-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00148-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00149-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00150-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00151-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00152-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00153-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00154-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00155-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00156-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00157-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00158-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00159-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00160-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00161-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00162-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00163-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00164-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
```

```
appy.parquet.crc
 extracting: random_data.parquet/.part-00165-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00166-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00167-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00168-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00169-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00170-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00171-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00172-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00173-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00174-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00175-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00176-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00177-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00178-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00179-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00180-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00181-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00182-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00183-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00184-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00185-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00186-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00187-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00188-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00189-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00190-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00191-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00192-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00193-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00194-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00195-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00196-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00197-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00198-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/.part-00199-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sn
appy.parquet.crc
 extracting: random_data.parquet/._SUCCESS.crc
```

```
  inflating: random_data.parquet/part-00000-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00001-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00002-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00003-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00004-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00005-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00006-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00007-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00008-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00009-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00010-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00011-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00012-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00013-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00014-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00015-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00016-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00017-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00018-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00019-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00020-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00021-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00022-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00023-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00024-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00025-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00026-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00027-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00028-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00029-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00030-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00031-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00032-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00033-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00034-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00035-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
```

```
  inflating: random_data.parquet/part-00036-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00037-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00038-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00039-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00040-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00041-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00042-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00043-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00044-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00045-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00046-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00047-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00048-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00049-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00050-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00051-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00052-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00053-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00054-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00055-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00056-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00057-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00058-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00059-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00060-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00061-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00062-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00063-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00064-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00065-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00066-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00067-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00068-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00069-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00070-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00071-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
```

```
  inflating: random_data.parquet/part-00072-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00073-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00074-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00075-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00076-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00077-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00078-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00079-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00080-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00081-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00082-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00083-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00084-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00085-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00086-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00087-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00088-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00089-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00090-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00091-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00092-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00093-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00094-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00095-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00096-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00097-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00098-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00099-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00100-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00101-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00102-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00103-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00104-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00105-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00106-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00107-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
```

```
  inflating: random_data.parquet/part-00108-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00109-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00110-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00111-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00112-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00113-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00114-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00115-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00116-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00117-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00118-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00119-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00120-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00121-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00122-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00123-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00124-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00125-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00126-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00127-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00128-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00129-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00130-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00131-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00132-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00133-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00134-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00135-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00136-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00137-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00138-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00139-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00140-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00141-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00142-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00143-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
```

```
  inflating: random_data.parquet/part-00144-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00145-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00146-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00147-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00148-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00149-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00150-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00151-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00152-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00153-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00154-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00155-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00156-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00157-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00158-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00159-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00160-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00161-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00162-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00163-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00164-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00165-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00166-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00167-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00168-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00169-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00170-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00171-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00172-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00173-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00174-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00175-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00176-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00177-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00178-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00179-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
```

```
  inflating: random_data.parquet/part-00180-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00181-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00182-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00183-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00184-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00185-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00186-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00187-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00188-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00189-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00190-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00191-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00192-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00193-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00194-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00195-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00196-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00197-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00198-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
  inflating: random_data.parquet/part-00199-4524816c-4134-4df9-9fe4-5fee65db2a4c-c000.sna
ppy.parquet
 extracting: random_data.parquet/_SUCCESS
```

SparkSession is created outside your function

In [ ]:

```
! pip install findspark
! pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publi
c/simple/
Requirement already satisfied: findspark in /usr/local/lib/python3.10/dist-packages (2.0.
1)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publi
c/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.10/dist-packages (3.4.0)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages
(from pyspark) (0.10.9.7)
```

In [5]:

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
import pyspark
from time import time

def init_spark(app_name: str):
  spark = SparkSession.builder.appName(app_name).getOrCreate()
  sc = spark.sparkContext
```

```
  return spark, sc

spark, sc = init_spark('hw2_kmeans')
```

Load samples points

In [6]:

```
data_df = spark.read.parquet("random_data.parquet").cache()
data_df.show(5)

# You can load the small sample for quick testing and reproducing results:

# sample_df = spark.read.option("header",True) \
#                      .option('inferSchema', True)\
#                      .csv('sample_data_84.csv')
# sample_df.show(5)
```

```
+-----+-----+-----+
|   _1|   _2|   _3|
+-----+-----+-----+
|8.186| 8.43|8.583|
|0.678|0.111|0.131|
|4.978|4.682|4.938|
|0.168|0.898|0.449|
|8.264|8.884| 8.88|
+-----+-----+-----+
only showing top 5 rows
```

Create initials centroids

In [7]:

```
init_centroids = spark.createDataFrame([[6.693, 7.782, 5.63],
                                        [3.744, 4.341, 7.225],
                                        [9.01, 7.8, 8.03],
                                        [2.134, 1.59, 1.93]]).cache()
init_centroids.show()
```

```
+-----+-----+-----+
|   _1|   _2|   _3|
+-----+-----+-----+
|6.693|7.782| 5.63|
|3.744|4.341|7.225|
| 9.01|  7.8| 8.03|
|2.134| 1.59| 1.93|
+-----+-----+-----+
```

In [14]:

```
from pyspark.ml.feature import VectorAssembler
import pyspark.sql.functions as F
from pyspark.ml.linalg import Vectors
import numpy as np
import sys
from pyspark.sql import types as T
from pyspark.ml.stat import Summarizer
from pyspark.sql import Row

# "Automating" the creation of vector out of a dataframe (that has the same schema as the
'data_df' dataframe)
col_names = [field.name for field in data_df.schema.fields]
assembler = VectorAssembler(inputCols=col_names, outputCol="point")

# Calculate distances between a given data point and each centroid, return the closest ce
ntroid's label.
def update_label(point, centroids):
```

```
  min_dist = sys.float_info.max
  for j in range(len(centroids)):
    # We don't have to sqrt the distances because the comparison relation still stays the
same,
    # it's still right to compare between the sqared distances - זהה סדר נשמר כעושים שורש
(על מספרים א' שליליים))
    dist = float(Vectors.squared_distance(point, centroids[j]))

    # Finding the closest centroid and updating the point's label to the centroid's corre
sponding label.
    if min_dist > dist:
      min_dist = dist
      label = j
  return int(label)

# Calculating the new centroid by the points assigned to its corresponding label accordin
g to average.
def update_centroid(data):
  new_centroid = []
  i = 0
  # Calculating the average of the points by calculating it separately on each vector com
ponent
  # and than combining them into a list according to the order of the components.
  for col in col_names:
    i += 1
    avg = data.select(col).agg(F.avg(data[col])).first()[0]
    new_centroid.append(avg)
  # Converting the centroid from type list to a vector.
  centroid_df = spark.createDataFrame([new_centroid], schema=col_names)
  return assembler.transform(centroid_df).select("point").first()[0]


def kmeans_fit(data: pyspark.sql.DataFrame,
               init: pyspark.sql.DataFrame,
               k: int = 4,
               max_iter: int = 10):
  """
  Inputs:
    data - a PySpark DataFrame that includes the data, as given to you from Moodle
    init - a PySpark DataFrame that holds the intial k centroids to be used in the algori
thm
    k - an integer - the amount of clusters in the algorithm
    max_iter - an integer - the maximum amount of iterations before terminating the algor
ithm

  Outputs:
    returns - centroids - a PySpark DataFrame that contains the final centroids from the
algorithm
  """

  # Adding a column named 'point' that will contain the values of all columns as a vector
in each row.
  data = assembler.transform(data_df).cache()
  data_df.unpersist()
  # Creating a list that contains all the centroids as vectors.
  centroids_list = assembler.transform(init_centroids).select("point").rdd.map(lambda x:
x[0]).collect()

  update_label_udf = F.udf(lambda point: update_label(point, centroids_list), T.IntegerT
ype())
  # Assigning two column so that one will contain the new updated labels and will contain
the old ones.
  # In each iteration (of updating labels) each column is being updated alternatively so
we could save the old labels easily.
  label_cols = ['label1', 'label2']
  labels = len(centroids_list)

  # Loop until we're exceeding the maximum number of iterations.
  for i in range(max_iter):
    j = i % 2
    # Updating the label of each point in the data.
    data = data.withColumn(label_cols[j], update_label_udf("point")).cache()
```

```
    # Check if the clusters have reached convergence (i.e., the assignments no longer cha
nges)
    if i > 0:
      count_label_changes = data.filter(data['label1'] != data['label2'])\
                                .count()
      if count_label_changes == 0:
        break

    temp_list = []
    # Re-calculating each cluster's centroid.
    for label in range(labels):
      cent_vec = update_centroid(data.filter(F.col(label_cols[j]) == label))
      temp_list.append(cent_vec)

    data.unpersist()
    # Updating the list of centroids to the newly calculated ones.
    centroids_list = temp_list

  # Converting the list of centroids to a dataframe, and returning it.
  return sc.parallelize([Row(centroids=centroid) for centroid in centroids_list]).toDF()
```

Test your function output and run time

In [15]:

```
start_time = time()
out = kmeans_fit(data_df, init_centroids).cache()
end_time = time()

print('Final results:')
out.show(truncate=False)
out.unpersist()
print(f'Total runtime: {end_time-start_time:.3f} seconds')
```

```
Final results:
+-------------------------------------------------------+
|centroids                                              |
+-------------------------------------------------------+
|[6.500257701999981,6.499862152000027,6.500300249000017] |
|[4.500187320000003,4.500350787999989,4.500139439999997] |
|[8.499745406999969,8.50008152299997,8.49965887499999]   |
|[1.5000080700000005,1.499990830999997,1.500135027500003]|
+-------------------------------------------------------+

Total runtime: 598.360 seconds
```