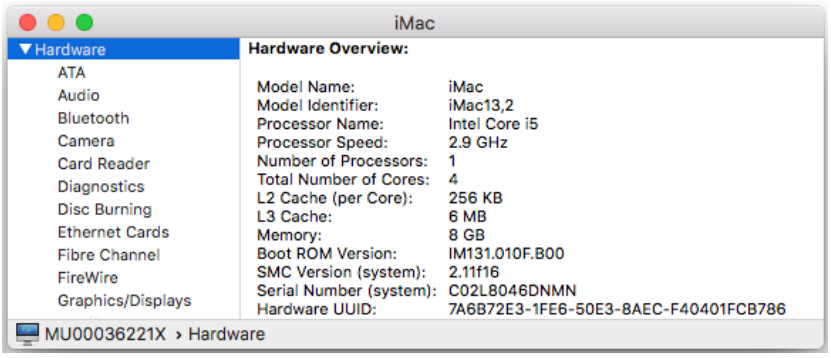
Question 1

Aditya and David are the first-year data science students with Monash University. They are discussing how parallel and distributed processing can help data scientists perform the computation faster. They would like your help to understand and get answers to the following questions:

1. Using the current processing resources, we can finish processing 1TB (one terabyte) of data in 1 hour. Recently the volume of data has increased to 2TB and the management has decided to double up the processing resources. Using the new processing resources, we can finish processing the 2TB in 60 minutes. Aditya wants to know **(1 + 1 = 2 Marks)**

    a. Is this speed-up or scale-up? Please explain your answer.
    b. Also, please explain what type of speed-up or scale-up is it (**linear, superlinear or sub-linear**)?

2. David is using his iMac desktop to do parallel query processing. The iMac has the following specifications:



   He wants to know what type of parallel database architecture is he using to do the parallel query processing. Please explain the reason for your answer. **(2 Marks)**

3. David read in the textbook that “Random unequal partitioning is sometimes inevitable in parallel search.” Please explain to him what is random unequal partitioning. **(2 Marks)**

4. Aditya now understands that skewness is the unevenness of workload and skewed workload distribution is generally undesirable. He found the figure below in the textbook that shows the skewed workload distribution. He wants to know **(1 + 2 =3 Marks)**
    a. Is the figure below **processing skew** or **data skew**? Please explain with reason.
    b. Is it possible to have an equal distribution of data? Please explain how.

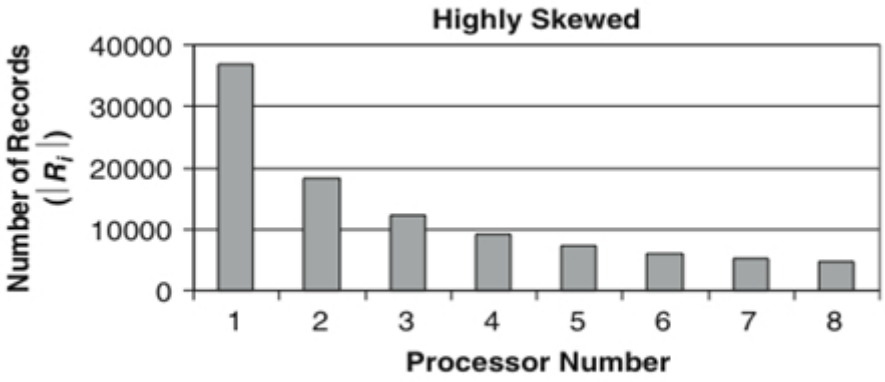


Figure 2.2 Highly skewed distribution

5. David was given a task to perform log analysis in the lab. The input data consisted of log messages of varying degrees of severity, along with some blank lines. He has to compute how many log messages appear at each level of severity. The contents of the “input.txt” file are shown below.

```
INFO This is a message with content
INFO This is some other content
(empty line)
INFO Here are more messages
WARN This is a warning
(empty line)
ERROR Something bad happened
WARN More details on the bad thing
INFO back to normal messages
```

The expected output of the operations is as below.

```
[(‘INFO’, 4), (‘WARN’, 2), (‘ERROR’, 1)]
```

However, he is not sure how to begin. Please explain to him assuming ‘sc’ as a SparkContext object. **(1 + 2= 3 Marks)**

a. What is an RDD?
b. How to read the “input.txt” file into an RDD?

Question 2

Petadata is an enterprise software company that develops and sells database analytics software subscriptions. The company provides three main services: business analytics, cloud products, and consulting. It operates in North and Latin America, Europe, and Australia.

Petadata is headquartered in Melbourne, Victoria, and has additional major Australian locations in Sydney and Adelaide, where its data center research and development is housed. Peter Liu has served as the company's president and chief executive officer since 2014. The company reported $2.8 billion in revenue, with a net income of $112 million, and 15,026 employees globally, as of March 15, 2020.

Chin is a recent graduate from Monash University and preparing for the job interview in Petadata. He needs your help to understand aspects of parallel processing especially parallel joins and parallel sort.

1. Using a more general notation, table R has |R| number of records, and table S has |S| number of records. The first step of ROJA is to redistribute the records from both tables according to hash/range partitioning. What is the **cost model** of the **Redistribution Step** of **ROJA**? **(4 marks)**

Symbol	**Description**
Data Parameters	
R	Size of table in bytes
R_i	Size of table fragment in bytes on processor i
|R|	Number of records in table R
$\|R_i\|$	Number of records in table R on processor i
Systems Parameters	
N	Number of processors
P	Page size
Time Unit Cost	
IO	Effective time to read a page from disk or write a page to disk
t_r	Time to read a record in the main memory
t_w	Time to write a record to the main memory
t_d	Time to compute destination
Communication Cost	
m_p	Message protocol cost per page
m_l	Message latency for one page

2. Given a data set D = {55; 30; 68; 39; 1; 4; 49; 90; 34; 76; 82; 56; 31; 25; 78; 56; 38; 32; 88; 9; 44; 98; 11; 70; 66; 89; 99; 22; 23; 26} and three processors, show step-by-step how the Parallel Redistribution Merge-All Sort works. **(5 Marks)**

   Assume random equal partitioning has been applied, where each processor has 10 records. The first processor will get the first 10 records, etc.

   Processor 1 = {55; 30; 68; 39; 1; 4; 49; 90; 34; 76}

   Processor 2= {82; 56; 31; 25; 78; 56; 38; 32; 88; 9}

   Processor 3 = {44; 98; 11; 70; 66; 89; 99; 22; 23; 26}

3. Chin was thinking of using internal sorting to perform the sort. However, he read on the internet that "**External Sorting** is different from **Internal Sorting**. Therefore, external sorting cannot use any of the Internal sorting methods". Is this statement True or False? Explain the reason as well. **(3 Marks)**

Question 3

2020 has been the year of Big Data – the year when big data and analytics made tremendous progress through innovative technologies, data-driven decision making and outcome-centric analytics. You are applying for the job as a Data Scientist. Mohammad is a senior lecturer and data scientist at Monash University, and a good friend of yours. He has prepared a list of questions regarding Apache Spark and Machine Learning to help you prepare for the job interview. Please answer the following questions.

1. In Apache Spark, machine learning pipelines provide a uniform set of high-level APIs built on top of DataFrames. It makes it easier to combine multiple algorithms into a single pipeline, or workflow. The key concepts introduced by the Pipelines API are DataFrame, Transformer, Estimator, Pipeline, and Parameter.
    a. What is Machine Learning and why should you use machine learning with Spark? **(2 Marks)**
    b. What is a Transformer and an Estimator? **(2 Marks)**

2. According to McKinsey study, 35% of what consumers purchase on Amazon and 75% of what they watch on Netflix is driven by machine learning–based product recommendations.
    a. Mohammad wants to know if you have understood how these recommendation systems work. So, please use the dataset below to recommend Top-2 movies to Mohammad. Please show all the calculations. **(4 Marks)**

Name	**StarTrek**	**StarWars**	**Superman**	**Batman**	**Hulk**
Mohammad	4	2	?	5	4
Paras	5	3	4	?	3
Huashun	3	**?**	4	4	3

    b. You are given a dataset “ratings” which contains movie ratings consisting of user, movie, rating and timestamp columns. The column names are *userId, movieId, rating and ts* respectively. Write a basic Machine Learning Program in PySpark to build the recommendation model and to make recommendation. Write the missing code snippets in the program given below. **(4 Marks)**

```
from pyspark.ml.recommendation import _____

Task #1: # split the dataset into training and test data (80% training and 20% test)
(trainingData, testData) = _____

Task #2: Build the recommendation model using ALS on the training data
# Use maxIter = 10, coldStartStrategy = “drop”

# make predictions
predictions = model.transform(testData)

Task #3: # Generate top 10 movie recommendations for each user
# Write code below
```

Question 4

StopHacking is a start-up incubated in Monash University to develop cloud service to detect and stop computer hackers. Although they have some rule-based service to identify certain hacks, they would like to add machine learning models which can integrate with their Spark cluster to process large amounts of data and detect any potential hacks. The dataset contains an “attack” column representing whether the request was an attack or not.

They hired you as the Lead *Data Scientist* and Peter (your intern) to investigate the open data from the Cyber Range Labs of UNSW Canberra and build a model based on the data to identify abnormal system behaviour.

Before proceeding with the development of ML models, Peter has some questions in mind that he would like your input on.

1. Peter is not sure whether this is a classification or a regression problem. Is this a classification or a regression problem? Briefly discuss when do we use classification and regression with examples. **(2 Marks)**

2. Upon investigation of the data, Peter has found that the data is imbalanced. Please suggest ways to handle an imbalanced dataset. **(2 Marks)**

3. You have prepared an estimator for the Decision Tree Model. Executing a Decision tree algorithm is a simple task. But, Peter still has some doubts. **(2 + 3 = 5 Marks)**
    a. How does a tree splitting take place? Explain in the context of the ID3 algorithm.
    b. The models perform great on the training data but generalize poorly to new instances. Peter is not sure what is happening. Can you explain what is happening and suggest two possible solutions.

4. What are False Positive(FP) and False Negative(FN) in a confusion matrix? Which value should we try to reduce in this scenario, discuss briefly? **(3 Marks)**

Question 5

Spectroscopy products developed at Divergent Technologies generate a lot of performance and diagnostic data. The data is typically stored locally on the controlling PC's hard disk drive and only analysed for the purpose of reviewing function and performance as a part of short term test requirements. Further analysis (such as trend analysis, predictive analytics, comparative studies, regression / correlation, etc.) is currently very challenging and is done manually on an as-needs basis.

You and Neha have been hired as summer interns to implement machine learning algorithms with the data generated by the spectroscopy products. These spectroscopy products have sensor arrays installed and it is anticipated that using ML techniques could prove extremely valuable that enable timely preventative maintenance of the sensors and / or responsive lower cost repairs. Ultimately, it may lead to the development of a sale-able product in this area, with potential use across the broader Divergent instrument portfolio.
You are working on streaming data from the sensors and Neha has some questions for you before she can develop the machine learning models.

1. The spectroscopy product has multiple sensors attached to it that measures different things for example light, gas and heat emission. Can you please explain two different methods that can be used to lower the granularity of the sensor arrays? **(4 Marks)**

2. There are three main sensors in the Spectroscopy products. So, Neha is planning to send the data using three Kafka producers using the same topic "spectroscopy_streams". The sensors are producing data as key value pairs in the format below and sent as bytes. **(4 Marks)**

   "gas": 125
   "light": 3298
   "heat": 78

   In the Apache Spark Streaming, the received data looks like below.

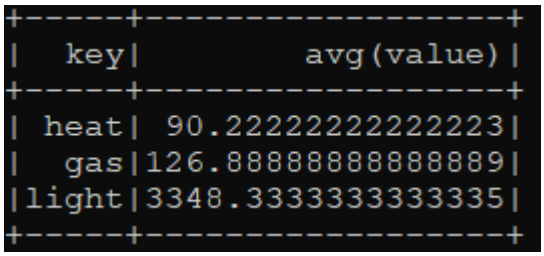
```
+----------------+-------------+--------------------+---------+------+--------------------+-------------+
|             key|        value|               topic|partition|offset|           timestamp|timestampType|
+----------------+-------------+--------------------+---------+------+--------------------+-------------+
|      [67 61 73]|   [31 34 30]|spectroscopy_streams|        0|   261|2020-10-18 00:06:...|            0|
|[6C 69 67 68 74]|[33 31 31 31]|spectroscopy_streams|        0|   262|2020-10-18 00:06:...|            0|
|   [68 65 61 74]|   [31 32 33]|spectroscopy_streams|        0|   263|2020-10-18 00:06:...|            0|
+----------------+-------------+--------------------+---------+------+--------------------+-------------+
```

   Please complete the code below for Apache Spark Streaming to find the average for each sensor every 10 seconds.

```
from pyspark.sql import SparkSession

spark = SparkSession. ...

Task #1: # Subscribe to the topic "spectroscopy_streams". The server is running on 192.168.0.10, port 9092.

Task #2:  Find the average for each sensor.

Task #3: # Start running the query that prints the running counts to the console every 10 seconds.
```

```
query.awaitTermination()
```

The output will be as shown in the example below.



```
+-----+------------------+
|  key|         avg(value)|
+-----+------------------+
| heat| 90.22222222222223|
|  gas|126.88888888888889|
|light|3348.3333333333335|
+-----+------------------+
```

3. Is the windowing method mentioned in the question time based window or tuple based window? Please explain. How can you enable time based overlapping sliding windows in Apache Spark Structured Streaming? **(4 Marks)**