Data Intensive Computing - Review Questions 3

Sri & Nik - 15th Sept, 2019

Question 1: DIfference between DataFrame and DataSet

A DateFrame has untyped rows, while DataSet has types rows.

When you rely on types, such as in map functions, it can be useful:

```
val peopleRDD = sc.parallelize(Array(Person("nik", 23, 0),
Person("sri", 11, 1)))
val peopleDS = peopleRDD.toDS
val collectedNames = collectedPeople.map {row => (row(0)}
```

Question 2: Find output and explain the code

Output

```
+-------+---+------------------+
|   name|age|           avg_age|
+-------+---+------------------+
|Michael| 15|              22.5|
|   Andy| 30|21.333333333333332|
| Justin| 19|20.333333333333332|
|   Andy| 12|16.666666666666668|
|    Jim| 19|14.333333333333334|
|   Andy| 12|              15.5|
+-------+---+------------------+
```

Explaination

```
val people = spark.read.format("json").load("people.json")
// Read the JSON
val windowSpec = Window.rowsBetween(-1, 1)
// For a given row, the window is the row itself and the row above and
below it, if any.
val avgAge = avg(col("age")).over(windowSpec)
// Average over the colomn age over the window of that row i.e r-1, r, r+1
people.select(col("name"), col("age"), avgAge.alias("avg_age")).show
// From people take colomns name, age and add avgAge computed earlier as
new colomn using alias and display the new content
```

Question 3: Difference between log-based broker systems and other broker systems

Log-based broker systems use logs to store the messages durably. Other systems use queues, which the broker deleted them once they're read from the queue.

Question 4: Compare windowing by processing time and event time + watermarks

A window (buffer) is used to collect messages and process them batch-wise. If a window is filled based on processing time, the system simply fills a window as the messages come in, and periodically cuts off the current window and starts of with a clean, empty one. If windowing is done based on event time, the system will fill in every window based on a predetermined starting and ending timestamp.

In the later, watermarks are special messages which are sent inside the data stream which denote the time. The time of the watermark tells the system that subsequent messages will not have been created earlier.

Question 5: Compare delivery guarantees

at-least-once mean that messages will be received 1 or more times. Care must be taken that when a message is received multiple times, it's not part of some calculation more than once.

exactly-once means that a message will only ever be consumed once. Here care must be taken that a message is not somehow skipped or lost in the computation.