Data Intensive Computing - Review Questions 3

Deadline: September 25, 2022

1. Briefly compare the `DataFrame` and `DataSet` in SparkSQL and via one example show when it is beneficial to use `DataSet` instead of `DataFrame`.

---

2. What will be the result of running the following code on the table `people.json`, shown below? Explain how each value in the final table is calculated.

```
val people = spark.read.format("json").load("people.json")
val windowSpec = Window.rowsBetween(-1, 1)
val avgAge = avg(col("age")).over(windowSpec)
people.select(col("name"), col("age"), avgAge.alias("avg_age")).show
```

people.json

```
{"name":"Michael", "age":15, "id":12}
{"name":"Andy", "age":30, "id":15}
{"name":"Justin", "age":19, "id":20}
{"name":"Andy", "age":12, "id":15}
{"name":"Jim", "age":19, "id":20}
{"name":"Andy", "age":12, "id":10}
```

---

3. What is the main difference between the log-based broker systems (such as Kafka), and the other broker systems.

---

4. Compare the windowing by processing time and the windowing by event time, and explain how watermarks help streaming processing systems to deal with late events.