Big Data Architectures
Winter 2024

Final Exam – Question 5

Question 5.1

Because the order of data coming from each sensor and across the two sensors is crucial, then the topic PressData should have 1 partition. Kafka can guarantee order within a partition but not across multiple one's. So, by having more than one partition would probably end up at risking of losing the order of messages. Now regarding the Consumer groups, in our scenario two applications are performing independent analytics processing, so therefore we must create 2 consumer groups, one for each application respectively. By having this setup we allow both application to independently consume all the messages from the topic PressData, while at the same time maintaining message order withing their processing logic.

Question 5.2

Because of the assumption that the banking application must be available 24/7/365, the property that must be sacrificed is Consistency, as per the CAP (Consistency, Availability, Partition tolerance) theorem states. In order to be able to ensure continuous availability even in the presence of network partitions or failures, the system could for even a small amount of time operate in a mode where not all nodes see the same data at the same time respectively, sacrificing that way strict consistency.

Question 5.3

When we refer to eventual consistency, we refer to a consistency model, in which its system guarantees that if no new updates are made to a given data item, eventually, all accesses to that aforementioned item will return back the same value. An example where such model could be located and it is acceptable is in a social media platform like Twitter, where seeing the most updated interactions like posts, comments etc. is not considered crucial at every moment. It is acceptable and tolerable from users to see these updates propagate and become consistent after a short period of time.

Question 5.4

Under the assumption that 25% of a program's execution is parallelizable, the maximum speedup (S) can be calculated using Amdahl's Law: $S=1/((1-P)+P/N)$, where P is the parallelizable portion, and N is the number of processors.

Plugging in the values: $S=1/((1-0.25)+0.25/3)=1/(0.75+0.25/3)\approx 1.2$

In practice, the achievement of this exact speedup might not be possible due to existence of overheads from parallel execution (e.g., synchronization, communication delays). Therefore, even though the theoretical speedup is about 1.2, the actual speedup might be less.

Question 5.5

The scenario of the dining philosophers or Deadlock, in our case is when three philosophers, each needing two chopsticks to eat, could all pick up their chopstick at their left at the same time, leaving none able to pick up their right chopstick because each philosopher's right chopstick is the left one of another philosopher. This results in a scenario called deadlock, where all philosophers hold one chopstick but can neither eat nor put down the chopstick for others to use. The reason that this scenario occurred is because of circular waiting, where each philosopher is waiting for a resource held by another. Regarding the solution, one approach to overcome this problem could be to to ensure that at least one philosopher picks up their right chopstick before their left chopstick, breaking with that mentality the circular wait condition that we aforementioned. Another solution to this problem is to allow a philosopher to pick up chopsticks, only in the case where both of his left and right chopsticks are available, ensuring that way that one philosopher will not end up holding onto one chopstick indefinitely if the other is not available.