12.05.2025 Week 12 exercises: Stream processing

Exercise 1:

(a) What is a Publish/Subscribe (pub/sub) system?

(b) Explain its basic functioning.

(c) What are the main advantages of pub/sub systems in comparison to direct messaging?

Exercise 2:

Concerning Kafka, answer the following questions:

(a) List and briefly explain the main roles in Kafka.

(b) What are topics and partitions? What ordering guarantees can a Kafka user expect?

(c) What is the purpose of the offset?

(d) What is the role of ZooKeeper?

(e) What are the responsibilities of Leaders and Followers?

Exercise 3:

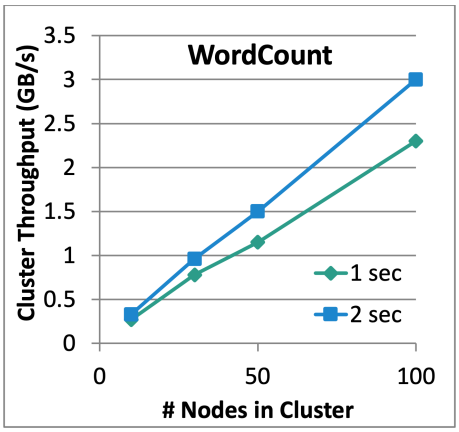
(a) What is is Windowing?

(b) Give a few examples on how to determine windows.

Exercise 4:

What is the difference between task and data parallelism?

Exercise 5:



The graph above shows the cluster throughput as a function of nodes in the cluster for a streaming word count example. Explain the impact of window size on the system latency. Explain why we achieve less throughput with a smaller window size.

Exercise 6:

True or False?

( ) In Topic-based pub/sub systems, a subscriber can specify filters on key/value attribute pairs of events.

( ) Kafka Producers manifest their interest in some topics and pull the corresponding data from brokers.

( ) Kafka users need to check for duplicates, as the system only guarantees at-least-once delivery.

( ) Continuous processing-based systems collect data in small groups and take action on each of them.

( ) A processing element (PE) operates on input tuples by applying a function on them and outputting other tuples.

( ) Pipelined parallelism consists of sequential stages that concurrently execute a computation on distinct data items.