**Spark Streaming / Kafka**

What is Spark Streaming?

- An extension of Spark that allows for the process of real-time data from various sources (such as Kafka)

What is Spark SQL?

* A spark module that allows for structured data processing

What is Structured Streaming?

* Structured streaming there is no batch. When new data is streamed in, it is appended to the result as opposed to being submitted in set batches of time. Data is received in an unbounded dataframe, while in unstructured streaming, data is reseaved in DStreams (batches of RDDs)

How does new data arriving in a stream get represented in Spark Streaming?

* Represented as RDD’s initially

In Structured Streaming?

* An unbounded table

What is Apache Kafka?

* Kafka is an event streaming platform used for working with data streams. It s a real-time messaging system.

What is Pub Sub?

* Pub Sub is a publish-subscribe architecture for messaging. The publisher writes a message to the topic and the subscriber listens for updates to a topic

Client Server?

* Kafka clients are created to read data from and write data to the kafka system
* Server provides services to the client
* Clients access data from the server

Messaging queues?

* Subscriber pulls a message from the end of a queue and the message gets removed

What are some advantages of pub sub? Why are those advantages especially relevant in large distributed applications?

* Messages are persisted in the cluster
* Asynchronous
* Scalable by simply adding consumers

What are events in Kafka?

* Events are records or messages. They contain a key and value and some optional metadata

What is a topic? A log?

* A topic is a category of messages in kafka
* A log is a segment of the data that provides a logical representation of a topic

Where do events come from in Kafka? Where do they go?

* Events come from a producer, stored in a topic, read by a consumer

What are the machines in a Kafka cluster called?

* Brokers

What are partitions in Kafka?

* The topics are split up into partitions that hold messages

How do partitions achieve HA and fault tolerance?

* Consumers grab data off of partitions. By having many partitions, we can have many consumers grab data, making the process scalable. If we have more consumers than partitions, some consumers will be “starved”, decreasing availability.
* Partitions are copied onto other brokers

What happens on machine failures in a Kafka cluster?

* https://medium.com/debutinfotech/achieving-fault-tolerance-with-kafka-a-detailed-explanation-a9828929d00d

What is Apache Zookeeper?

* Zookeeper manages the brokers in a kafka cluster

What’s the difference between pulling and pushing messages/events in a pub sub architecture? Which does Kafka use?

* Kafka is pull-based. In a pull based design, consumers pull date from the brokers. In a push based design, data is pushed onto the consumers by the broker. Pull-based allows for scalability as Kafka can have many consumers pulling data at different speeds

How long does Kafka store messages by default? What are our configuration options?

* Default 2 weeks
* log.retention

Does Kafka guarantee that it will preserve the order of produced events? What are the caveats?

* Kafka preserves order within a partition, not across partitions

How can we make sure that Kafka preserves the order of events where it matters?

* Using one partition for that event