Kafka and Spark Streaming

PRESENTER

PRAJOD VETTIYATTIL

ARCHITECT

OPEN SOURCE, BIG DATA

LOCATION

SPARK MEETUP, APIJEE,

BANGALORE

DATE

1 AUGUST 2015

Agenda

- Kafka
  - Concepts
  - How to use
  - When to use
- Spark Steaming
  - Micro batches
  - Time window API
  - Reliability

Kafka

Kafka concepts

- Message Oriented Middleware
- Brokers
- Topics
- Producers
- Consumers

Kafka broker

- Receive messages from producers
- Persist messages to log files
- Retrieve messages for consumers
- Zookeeper
  - Coordination
  - State management
  - Cluster configuration management
  - Storm, HBase

Topics

- Topics and Queues
- Consumers and Consumer groups
- Consumer offset
- Partitions
- One log for each partition

Producers

- Send message to topics
- Specify the partition, use custom partitioner or send to a random partition
- Zookeeper reference is needed
- Messages are sent to the kafka leader in a cluster

Consumers

- Consume messages from kafka brokers
- Simple API
- High Level API
- Read from any position
- Offset: position of message in a partition
- Consumers are mapped to partitions

Spark streaming

Micro batches

- Micro batch: A set of messages in a time window
- Discretized streams
- Receiver thread
- Processing thread
- Access to Spark platform API(GraphX, MLlib)

API

- Receivers: Kafka, Flume, Kinesis, Twitter, Socket, ZeroMQ
- Time window API: function performs once per batch
- Receiver API
- Direct API

Reliability

- Exaclty once semantics with HDFS
- Atleast once semantics with Receivers
- Checkpointing

References

- Kafka
  - http://kafka.apache.org/documentation.html
  - Install and run an example http://kafka.apache.org/documentation.html#quickstart
  - https://github.com/abhioncbr/Kafka-Message-Server/wiki/Apache-of-Kafka-Architecture-(As-per-Apache-Kafka-0.8.0-Dcoumentation)
- Spark Streaming
  - http://spark.apache.org/docs/latest/streaming-programming-guide.html
- Other references
  - http://www.slideshare.net/spark-project/deep-divewithsparkstreaming-tathagatadassparkmeetup20130617
  - http://www.slideshare.net/prajods/apache-spark-the-next-gen-toolset-for-big-data-processing