Big Data Technologies

Agenda

- Apache Kafka

Spark Kafka Integration

- Refer slides
- Spark dataframes are divided into the partitions. When dataframe created from kafka, one dataframe partition is mapped to one kafka partition.
- Spark dataframe partitions are on workers i.e. each worker will read corresponding partition from kafka topic partition. When a streaming query is executed, a consumer is created.
- one query(job) = one consumer group --> each group will have consumers equal to number of workers used to read from kafka.
- If two different jobs are processing two different kafka topics, number of dataframe partitions in each job is equal to number of kafka partitions in that topic. All tasks (threads) processing the partitions corresponding to one topic will be in one consumer group.
- If two different jobs are processing same kafka topic (for doing different processing), still there will be two different consumer groups. Each group will contain number of consumers equal to number of dataframe partitions (which in turn equal to number of kafka partitions). Note that data will be read twice from each kafka partition to create two different dataframe partitions (one for each job).

Event time processing

- Time at which event is generated at source, is "event time".
- Can process out-of-order data.
- Watermark feature is used to define for how much time late data should be considered (how much time should wait before processing data).

```
# assuming than values coming from kafka source (no key specified)
# values are json {"sensor": "LDR_B", "reading": 879, "time": "2024-01-10 10:33:10"}
result = data\
    .selectExpr("CAST(value AS STRING) val")\
    .selectExpr("FROM_JSON(val, 'sensor STRING, reading INT, time STRING') v")\
    .selectExpr("v.sensor", "v.reading", "CAST(v.time AS TIMESTAMP) AS time")\
    .withWatermark("time", "30 seconds")\
    .groupby("sensor", window('time', windowDuration='20 seconds', slideDuration='5
seconds')).avg("reading")\
```

Assignments

1. Load emp and dept data into spark dataframes. Plot various graphs like bar chart, pie chart, etc.
   - Deptwise total salary
   - Number of emps per Dept per Job
   - Salary proportion per emp (out of total sal)
2. Load a ML dataset. Perform EDA.