Structured Streaming – Basic Session 2

Segment - 01

Session Introduction

SESSION OVERVIEW

- What is Structured Streaming?
- Coding Lab
- Triggers and Output Modes
- Working with Structured Streams (With Coding Lab)
- Transformations and Aggregations (With Coding Lab)
- Joins with Streams
- Coding Lab for Joins
- Practice Coding Problems

Segment - 02

What is Structured Streaming?

WHY STRUCTURED STREAMING?

- High level API
- Ease of development
- Compatibility with other Spark APIs
- Spark Optimizations built in

KEY FUNDAMENTALS

- Spark principles stay in place
  - Lazy evaluation
  - Transformations
  - Actions
- Inputs
  - Streaming systems – Kafka, Flume
  - File systems – S3
  - Sockets
- Outputs
  - Databases
  - Input systems

CODE FLOW

- Create SparkSession
  - Entry point for a structured streaming job
- Read From Source
  - Socket/ Kafka/ File etc.
  - Read stream
  - Return DataFrame
- Perform Transformations
  - Create one of more DataFrames
- Start – Action
- AwaitTermination
  - Wait for the stream to finish

Segment - 03

Coding Lab

Segment - 04

Triggers and Output Modes

KEY FUNDAMENTALS

- Output Modes – What gets written to sink
  - Append = New records
  - Update = Modified records
  - Complete = Everything
- Restrictions on Output Modes
  - No aggregation => Update = Append
  - Append/ Update not allowed on Aggregations without Watermarks

KEY FUNDAMENTALS

- Triggers -> When new data gets processed in the stream
  - Default = When a new micro batch comes up
  - Once = A single micro batch
  - Processing time = Scheduler
  - Continuous = Each record level

Segment - 05

Coding Lab

Segment - 06

Transformations and Aggregations

KEY APIs

- PySpark.SQL
  - SQL functionalities of Spark
- Select
- SelectExpr
  - Any SQL-like statement/ expression
  - Takes a String as an argument

TRANSFORMATIONS & AGGREGATIONS

- Filter/ Where - To filter out some elements from the RDD which does not meet the criteria defined in the lambda expression
- As/ Aliasing – To make the output more readable by giving a different name aka aliasing
- GroupBy – Shuffle and group the data accordingly
- Aggregations
  - Min/ Max/ Avg/ Sum etc.

Coding Lab

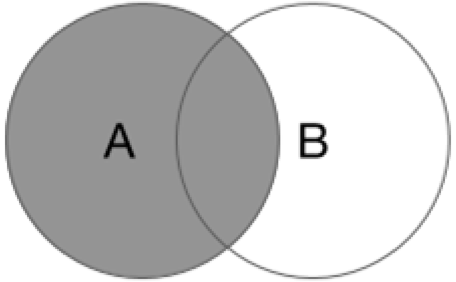
Segment - 07

Joins With Streams

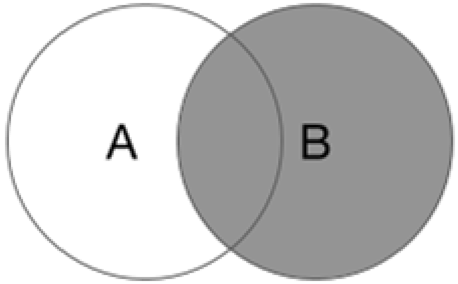
JOINS

- Similar to SQL Joins
  - Assume each streaming DF as a table
  - Stream DFs can join with other Stream DFs or Static DFs in the same way
- Inner Join
- Outer Join
  - Left
  - Right
  - Full

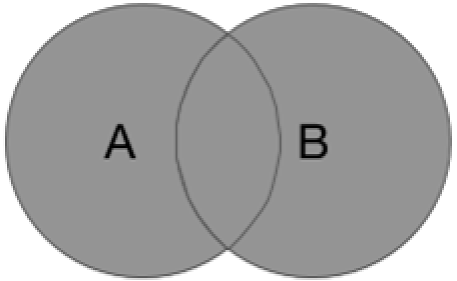
JOINS



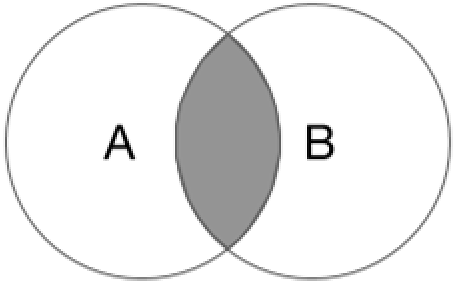
Left outer join



Right outer join



Full outer join



Inner join

JOINS

- Restrictions on Outer Joins
  - Stream - Stream Outer Joins only with Watermarks
  - Stream – Static Right Outer/ Full Outer Join not Permitted
  - Vice versa
  - Why?
  - Stream – Stream Full Outer Join not permitted
- Output Mode
  - Only Append supported for Stream – Stream Joins

Segment - 08

Coding Lab

Segment - 10

Session Summary

SESSION SUMMARY

- What is Structured Streaming?
- Key Fundamentals
- Flow of Code
- Source and Sinks
- Output Modes
- Triggers
- Transformations and Aggregations
- Joins
- Static + Stream
- Restrictions