Introduction to Spark Streaming - Session 1

Segment - 01

Module Introduction

MODULE INTRODUCTION

01 Introduction to Spark Streaming

02 Structured Streaming Basic

03 Structured Streaming Advanced

04 Industry Examples

Segment - 02

Session Introduction

SESSION OVERVIEW

- Streaming
- Differences between Streaming and Micro-Batching
- Spark Streaming

Segment - 03

Streaming

DATA STREAMS

- Continuous inflow of Data = Stream of Data
  - Amazon.com's Log Data
  - Live video
  - IoT devices' data
- No discrete start or end of data
- Usually, high-volume data
- (Near) Real-time processing and response

WHY STREAMS

- Amazon.com’s log data
  - Scaling up/ down
  - HW/ Data center incident response
- Live Video
  - HW scaling
  - Real-time reactions
  - Analytics
- IoT Data
  - Factory response based on machine data
- Fraud Detection
  - Real-time detection
  - Stop transaction completion

STREAMING DATA ARCHITECTURE

- Framework built for ingesting and processing streams of data
- Multiple Components
  - Stream consumer
  - Data persistence
  - Processing/ Transformation
  - Analytics/ BI
- Consume and act on data immediately

Segment - 04

Differences Between Streaming and Micro-Batching

BATCH PROCESSING

- Each execution processes a batch
  - Based on a set time window
  - Hourly/ Daily/ Weekly/ Monthly etc
- Capability of producing high volumes of data
- Substantial latency for the BI layer
- What happens if you keep reducing the batch size?

MICROBATCHES OR STREAMING

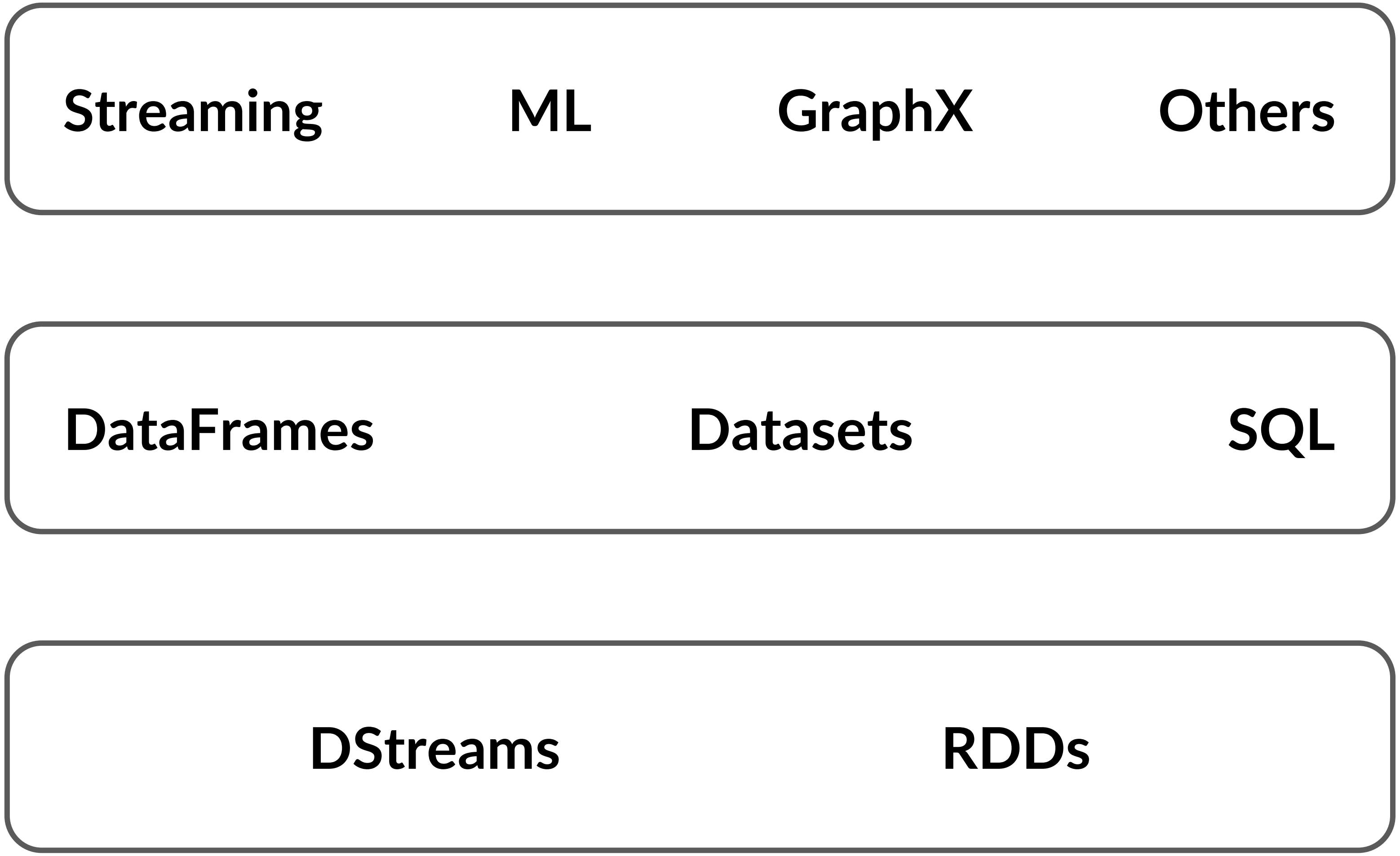
- Days -> Hours -> Minutes
- Near real-time data availability
- Spark Streaming runs on micro batches, with the trigger set to 0, so the data is read continuously
- Practically, same as streaming

Segment - 05

Spark Streaming

SPARK ARCHITECTURE

- Applications
- High-Level APIs (Structured APIs)
- Low-Level APIs



Streaming ML GraphX Others

DataFrames Datasets SQL

DStreams RDDs

SPARK STREAMING

- Declarative API – Spark decides where to run what
- Spark Streaming runs on micro batches
- Structured Streaming
  - High-level API
- DStreams
  - Low-level API

STRUCTURED STREAMING VS DSTREAMS

- Structured Streaming
  - Similar to Batch Spark processing
  - DataFrames + SQL
  - Spark optimizations automatically picked up
  - Exactly once guarantee
- DStreams
  - Collection of RDDs
  - Processing late arriving data problematic
  - Inconsistencies with RDDs, datasets APIs

Segment - 06

Session Summary

SESSION SUMMARY

- Batch vs Stream Processing
- Stream Processing Architecture
- Spark Streaming = Micro-Batches
- Spark Streaming APIs
- Structured Streaming vs DStreams