McGill | School of Continuing Studies | École d'éducation permanente

YCBS-257 - Data at Scale

Workshop 10

Part 1

Real-Time Data Analysis

Spark Structured Streaming Basics

Overview

Apache Spark is one of the most widely adopted engines for large-scale data processing and analytics. Traditionally, Spark has been used to process **static datasets** - data at rest. However, with the introduction of the Spark Streaming module, it became possible to process **live data streams** in real time, enabling immediate insights as data arrives.

Spark Structured Streaming, introduced in Spark 2.2+, is a modern, high-level API for real-time stream processing. It builds on Spark's structured APIs and provides a more intuitive, powerful, and fault-tolerant framework compared to the older Spark Streaming engine, which is now considered legacy and no longer maintained.

With Structured Streaming, you can use the **same operations** that you would in batch processing - such as filters, aggregations, joins, and windowing - and run them in streaming mode with minimal or no changes to your code. This significantly reduces complexity and lowers the barrier to entry for developing robust real-time data applications.

Official Documentation

For more information, refer to the official Spark site:
https://spark.apache.org/

Note on Environment

In production environments, Spark Structured Streaming applications typically run on dedicated servers with sufficient resources to efficiently **ingest and process data streams in real time**.

In this workshop, however, we will use **Apache Zeppelin** to develop and run our streaming applications. While Zeppelin is not optimized for real-time streaming workloads, it is **suitable for learning and experimentation**, making it a good choice for gaining hands-on experience with the core concepts of Structured Streaming.

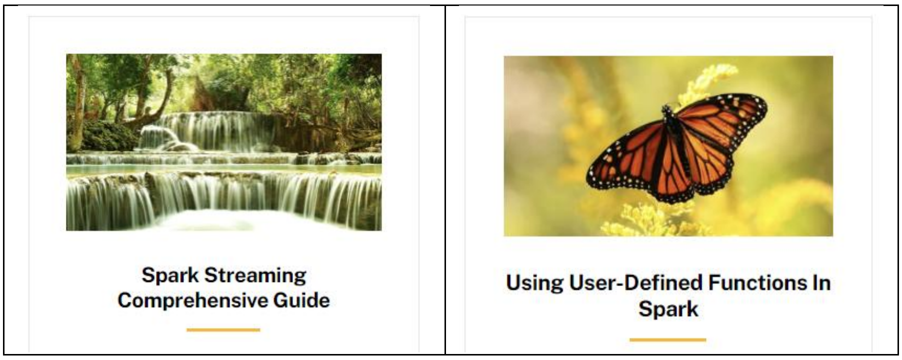
Spark Structured Streaming

Tutorials

Learning Activities

These tutorials are designed to help you **get started with Apache Spark Structured Streaming**. You will learn how to use **Structured Streaming in combination with the Spark SQL DataFrame API** to read data from streaming sources, process it in real time, and persist the results for further analysis.

To begin, navigate to the **Tutorials** page in the **DDP sandbox** and explore the following modules:



Spark Streaming Comprehensive Guide

Using User-Defined Functions In Spark