Apache Storm – Course Outline

1 Duration

16 Hours (4 half days)

2 Objectives

This course provides good coverage of Apache Storm. At end of this course, participants should

- Master the fundamental concepts and the architecture of Apache Storm
- Plan installation and configuration with Apache Storm
- Grasp concepts such as Ingesting and processing of real-time events with Storm
- Understand fundamentals of Trident extension to Apache Storm
- Gain thorough understanding of Grouping & Data Insertion in Apache Storm

3 Audience

Developers and Data Engineers who are looking to build real time stream processing pipeline.

4 Pre-requisite

- General knowledge on messaging systems and concepts
- Knowledge on programming (preferably Java)
- Basic knowledge of Linux or Unix based systems
- Basic knowledge of data processing

5 Hardware & Network Requirements

- Desktop/Laptop with minimum 16GB RAM
- Proxy Free High speed internet connection (minimum 5 Mbps)

6 Software Requirements

- Windows or Linux or Mac
- JDK 11+
- Putty / SSH Client
- Eclipse latest version
- Docker Desktop / Engine

7 Outline

Day 1

Module-1: Introduction to Storm

- Why Storm?
- What Storm does?
- How Storm works?
- Overview of a modern data architecture
- Conceptual overview of Storm
    - Topologies
    - Streams
    - Spouts
    - Bolts
    - Stream groupings
    - Reliability
    - Tasks
    - Workers
- Overview of Scheduler
- Setting up a Development Environment
    - Setup Storm
    - Local vs remote mode
    - Installing Storm locally
    - Starting and stopping topologies on remote cluster
    - Creating a new Storm Project
- Stream Processing Overview
- Stream Processing Usecases

Day 2

Module-2: Storm Architecture and Programming

- Create Storm topologies and deploying them on Storm cluster
    - Components of a Storm cluster
    - Topologies
    - Streams
    - Data model
    - A simple topology
    - Running topology in local mode
    - Stream groupings
    - Multi-language support using Bolts in other languages
    - Overview of Guaranteeing message processing
    - Overview of Transactional topologies
    - Overview of Distributed RPC
- Storm Configuration
    - Setting up Multi-node Storm Cluster
- Spouts & Bolts

  - Types of Spouts
  - Structure of Spout
  - Structure of Bolt
  - Stream Groupings
  - Reliable processing in Storm
  - Ack and Fail
  - Ack Timeout

Day 3

Module-3: Real-time Data Processing Pipeline with Storm

- Topology lifecycle
- Data ingestion in Storm
- Real time data ingestion
- Topology connecting Spout and Bolt
- Setting up a Storm Cluster
  - Set up a Zookeeper cluster
  - Install dependencies on Nimbus and worker machines
  - Download and extract a Storm release to Nimbus and worker machines
  - Configure in storm.yaml
  - Launch daemons using storm script and a supervisor
- Kafka integration with Storm
- Create data processing pipeline with sample dataset with Kafka integration

Day 4

Module-4: Intro to Storm SQL and Trident

- Storm SQL Overview
  - Streaming from and to external data sources
  - Filtering tuples
  - Projections
  - User defined function
- Overview of Trident
  - DRPC topology
  - Fields and tuples
  - State
  - Execution of Trident topologies
  - Trident spouts
  - Trident Resource Aware Scheduler

Module-5: Intro to Heron

- Heron Overview
- Heron Architecture
- Storm vs Heron
- Heron Benefits and Limitations
- Heron Usecases