Training Course on Real-time Data Analytics and Stream Processing (Kafka and Flink/Spark Streaming)

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

In today's hyper-connected, data-driven landscape, the ability to extract immediate, actionable insights from constantly flowing information is paramount. Training Course on Real-time Data Analytics and Stream Processing (Kafka & Flink/Spark Streaming) empowers professionals to master the cutting-edge technologies that power live data pipelines. We delve into the intricacies of **Apache Kafka** for robust data ingestion and messaging, and then explore **Apache Flink** and **Apache Spark Streaming** for sophisticated, scalable stream processing. Participants will gain hands-on expertise in building fault-tolerant, high-throughput architectures crucial for modern enterprises seeking to capitalize on **instantaneous data intelligence** and drive **real-time decision-making**.

This comprehensive program moves beyond theoretical concepts to practical implementation, focusing on architecting and deploying **end-to-end streaming solutions**. From **event-driven architectures** to **stateful stream processing** and **real-time machine learning**, attendees will learn to transform raw data streams into valuable business insights. Through a blend of expert-led instruction, interactive labs, and real-world case studies, this course cultivates the skills necessary to design, build, and optimize **live data pipelines** that unlock competitive advantages in diverse industries, including finance, IoT, e-commerce, and cybersecurity, ensuring organizations can respond to opportunities and threats with unparalleled agility.

Programme Curriculum

Training Course on Real-time Data Analytics and Stream Processing (Kafka & Flink/Spark Streaming)

Introduction

In today's hyper-connected, data-driven landscape, the ability to extract immediate, actionable insights from constantly flowing information is paramount. Training Course on Real-time Data Analytics and Stream Processing (Kafka & Flink/Spark Streaming) empowers professionals to master the cutting-edge technologies that power live data pipelines. We delve into the intricacies of **Apache Kafka** for robust data ingestion and messaging, and then explore **Apache Flink** and **Apache Spark Streaming** for sophisticated, scalable stream processing. Participants will gain hands-on expertise in building fault-tolerant, high-throughput architectures crucial for modern enterprises seeking to capitalize on **instantaneous data intelligence** and drive **real-time decision-making**.

This comprehensive program moves beyond theoretical concepts to practical implementation, focusing on architecting and deploying **end-to-end streaming solutions**. From **event-driven architectures** to **stateful stream processing** and **real-time machine learning**, attendees will learn to transform raw data streams into valuable business insights. Through a blend of expert-led instruction, interactive labs, and real-world case studies, this course cultivates the skills necessary to design, build, and optimize **live data pipelines** that unlock competitive advantages in diverse industries, including finance, IoT, e-commerce, and cybersecurity, ensuring organizations can respond to opportunities and threats with unparalleled agility.

Course Duration

10 days

Course Objectives

1. Understand and implement highly scalable, fault-tolerant data ingestion strategies using Apache Kafka for high-volume event streams.
2. Architect robust systems leveraging Kafka as the central nervous system for asynchronous communication and decoupled microservices.
3. Grasp core concepts of stream processing, including windowing, state management, and event-time vs. processing-time semantics.
4. Develop and deploy complex, stateful stream processing applications using Apache Flink's powerful DataStream API for low-latency analytics.
5. Implement efficient, scalable real-time data pipelines with Apache Spark Structured Streaming, utilizing DataFrame/Dataset APIs.
6. Design, implement, and optimize complete streaming data pipelines from source to sink, integrating Kafka, Flink/Spark, and downstream systems.
7. Apply advanced techniques for transforming, filtering, aggregating, and enriching streaming data in motion.
8. Design and build resilient streaming architectures with built-in fault tolerance, data durability, and disaster recovery mechanisms.
9. Create interactive, dynamic real-time dashboards and visualizations to expose immediate business insights.
10. Deploy and manage machine learning models for continuous, real-time predictions and anomaly detection on streaming data.
11. Optimize streaming applications and infrastructure for maximum throughput, minimal latency, and efficient resource utilization.
12. Implement effective monitoring, alerting, and debugging strategies for real-time data pipelines in production environments.
13. Understand deployment strategies for Kafka, Flink, and Spark Streaming on major cloud platforms (AWS, Azure, GCP) for scalable and managed solutions.

Organizational Benefits

- Enable immediate responses to market shifts, customer behavior, and operational events, leading to a significant competitive advantage.
- Provide real-time personalization, proactive customer support, and instant recommendations based on live user interactions.
- Optimize business processes, detect anomalies, and predict maintenance needs in real-time, reducing downtime and costs.
- Instantly identify and mitigate fraudulent activities and security breaches as they occur, minimizing financial losses and risks.
- Foster the development of innovative, data-driven products and services that leverage instantaneous insights.
- Build highly scalable and flexible data architectures that can adapt to evolving business needs and increasing data volumes.
- Reduce storage and processing costs by processing data in motion rather than relying solely on batch processing and large data warehouses.
- Cultivate a culture of data-driven decision-making throughout the organization by providing timely, relevant insights to all stakeholders.

Target Audience

1. Data Engineers.
2. Software Developers.
3. Data Architects.
4. Big Data Developers.
5. DevOps Engineers
6. Data Scientists.
7. Technical Leads & Managers
8. Solutions Architects.

Course Outline

Module 1: Introduction to Real-time Data & Streaming Fundamentals

- What is Real-time Data Analytics? Definition, benefits, and use cases.
- Batch vs. Stream Processing: Key differences, advantages, and limitations.
- Core Concepts: Events, streams, producers, consumers, and data pipelines.
- Architectural Patterns for Streaming Systems: Lambda, Kappa, and event-driven architectures.
- Challenges in Real-time Processing: Data velocity, volume, variety, and veracity.
- *Case Study:* Real-time fraud detection in financial transactions (e.g., credit card fraud).

Module 2: Apache Kafka: The Distributed Streaming Platform

- Kafka Architecture: Brokers, topics, partitions, offsets, and consumer groups.
- Kafka Producers: Writing data to Kafka topics efficiently and reliably.
- Kafka Consumers: Reading and processing data from Kafka topics, consumer offsets.
- Kafka Connect: Integrating Kafka with various data sources and sinks (databases, S3, etc.).
- Kafka Administration & Operations: Setting up clusters, monitoring, and basic troubleshooting.
- *Case Study:* Building a centralized logging and event hub for a large e-commerce platform.

Module 3: Advanced Kafka Concepts & Best Practices

- Schema Registry & Avro/Protobuf: Managing data schemas for evolution and compatibility.
- Kafka Security: Authentication, authorization (ACLs), and encryption.

- Kafka Streams API (Overview): Lightweight stream processing within Kafka applications.
- Kafka KSQL (Overview): SQL-like interface for real-time querying of Kafka topics.
- Partitioning Strategies & Performance Tuning: Optimizing Kafka for high throughput and low latency.
- *Case Study:* Real-time user activity tracking for website personalization using Kafka events.

Module 4: Introduction to Apache Flink: Stream Processing Powerhouse

- Flink Architecture: JobManagers, TaskManagers, slots, and distributed execution.
- Flink DataStream API Fundamentals: Transformations (map, filter, flatMap).
- Stateful Stream Processing: Managing and accessing state in Flink applications.
- Time Concepts: Event time, processing time, and ingestion time.
- Fault Tolerance & Checkpointing: Ensuring data consistency and recovery in Flink.
- *Case Study:* Real-time sensor data aggregation and anomaly detection in IoT.

Module 5: Advanced Apache Flink: Windows & Joins

- Windowing Operations: Tumbling, sliding, session, and global windows.
- Event Time Processing & Watermarks: Handling out-of-order events.
- Joins in Flink: Stream-to-stream joins and stream-to-table joins.
- Connectors & Integrations: Connecting Flink with Kafka, databases, and other systems.
- Deployment & Operations: Deploying Flink applications on YARN, Kubernetes, or standalone.
- *Case Study:* Real-time stock market analysis by joining multiple data feeds.

Module 6: Apache Spark Streaming & Structured Streaming

- Spark Streaming DStreams (Legacy Overview): Micro-batch processing concept.
- Spark Structured Streaming Fundamentals: Continuous processing on unbounded tables.
- DataFrame/Dataset API for Streaming: Using familiar Spark APIs for stream processing.
- Stateful Operations in Structured Streaming: Aggregations, joins, and watermarks.
- Fault Tolerance & Exactly-Once Semantics in Structured Streaming.
- *Case Study:* Real-time sentiment analysis of social media feeds with Spark Structured Streaming.

Module 7: Choosing the Right Stream Processor: Flink vs. Spark Streaming

- Feature Comparison: Latency, state management, fault tolerance, API paradigms.
- Use Case Suitability: When to choose Flink, when to choose Spark Streaming.
- Performance Benchmarking: Understanding metrics and optimizing for specific workloads.
- Ecosystems & Community Support: Comparing the broader tools and resources available.
- Hybrid Approaches: Combining Kafka Streams, Flink, and Spark for complex scenarios.
- *Case Study:* Designing a real-time analytics platform for a gaming company, deciding between Flink and Spark based on game event characteristics.

Module 8: Building End-to-End Real-time Data Pipelines

- Pipeline Design Principles: Modularity, scalability, and maintainability.
- Data Ingestion Strategies: From various sources (logs, IoT sensors, databases).
- Data Transformation & Enrichment: Cleaning, normalizing, and augmenting data in transit.
- Data Storage & Serving Layers: Real-time databases (e.g., Cassandra, MongoDB, ClickHouse).
- Integration with Downstream Systems: Dashboards, alerting systems, and applications.

- *Case Study:* Constructing a live supply chain monitoring system from sensor data to an operations dashboard.

Module 9: Real-time Analytics & Business Intelligence

- Real-time Dashboards: Designing effective visualizations with tools like Grafana, Kibana, Tableau.
- Key Performance Indicators (KPIs) in Real-time: Identifying and tracking critical metrics.
- Anomaly Detection: Techniques for identifying unusual patterns in live data streams.
- Alerting Mechanisms: Setting up notifications for critical events or threshold breaches.
- Ad-hoc Querying on Streaming Data: Leveraging tools like KSQL or Spark SQL for real-time insights.
- *Case Study:* Real-time monitoring of network traffic for security threats and performance issues.

Module 10: Real-time Machine Learning with Streaming Data

- Overview of Streaming ML: Challenges and opportunities.
- Online Learning: Continuously updating models with new data.
- Real-time Feature Engineering: Preparing data for ML models on the fly.
- Deploying ML Models for Real-time Inference: Integrating with Flink or Spark.
- Use Cases: Predictive maintenance, recommendation engines, fraud detection.
- *Case Study:* Building a real-time recommendation engine for an online streaming service.

Module 11: Deployment & Operations of Streaming Systems

- Containerization (Docker) & Orchestration (Kubernetes) for Kafka, Flink, Spark.
- Monitoring & Logging: Tools and best practices for observing pipeline health.
- Alerting & Incident Response: Setting up robust alert systems.
- Performance Tuning & Optimization: Strategies for maximizing throughput and minimizing latency.
- Disaster Recovery & Business Continuity for Streaming Architectures.
- *Case Study:* Deploying and managing a production-grade real-time advertising bidding platform.

Module 12: Cloud-Native Real-time Analytics

- Managed Kafka Services: Confluent Cloud, AWS MSK, Azure Event Hubs.
- Managed Flink/Spark Services: AWS Kinesis Data Analytics, Google Cloud Dataflow, Databricks.
- Serverless Streaming Architectures: Leveraging cloud functions and managed services.
- Cost Optimization in Cloud Environments for Streaming Workloads.
- Security and Compliance in Cloud-based Streaming.
- *Case Study:* Migrating an on-premise real-time data pipeline to a fully managed cloud solution.

Module 13: Data Governance & Quality in Streaming

- Data Lineage & Metadata Management for real-time streams.
- Schema Evolution & Compatibility in dynamic data environments.
- Data Validation & Quality Checks: Ensuring accuracy and completeness of streaming data.
- Privacy & Compliance (GDPR, CCPA): Handling sensitive data in motion.
- Auditing & Logging for traceability and accountability.

- *Case Study:* Ensuring data quality and compliance for a real-time patient monitoring system in healthcare.

Module 14: Advanced Stream Processing Patterns & Architectures

- Complex Event Processing (CEP): Detecting patterns across multiple events.
- Backpressure Handling: Managing uneven data flow in streaming pipelines.
- Exactly-Once Processing Semantics: Guarantees for data consistency.
- Distributed Transactions in Streaming Contexts.
- Microservices and Event Sourcing with Kafka.
- *Case Study:* Implementing real-time supply chain optimization using CEP to react to disruptions.

Module 15: Future Trends & Emerging Technologies

- Real-time Data Mesh & Data Products.
- Stream Processing with AI/ML integration.
- Edge Computing & Stream Processing at the Edge.
- Next-generation Streaming Databases (e.g., RisingWave, Materialize).
- Impact of Serverless and Function-as-a-Service on streaming.
- *Case Study:* Exploring the potential of real-time analytics in autonomous vehicle data processing.

Training Methodology

Our training methodology is highly interactive and hands-on, designed to ensure practical skill acquisition and deep understanding.

- **Instructor-Led Sessions:** Expert-led discussions, theoretical explanations, and concept deep-dives.
- **Live Coding Demonstrations:** Step-by-step building of real-time data pipelines and applications.
- **Hands-on Labs & Exercises:** Extensive practical sessions where participants implement concepts using real-world datasets and tools (Kafka, Flink, Spark, etc.).
- **Case Studies & Real-world Scenarios:** Analysis and discussion of industry-specific applications of real-time data analytics.
- **Group Discussions & Problem Solving:** Collaborative sessions to tackle complex challenges and share insights.
- **Q&A and Interactive Troubleshooting:** Dedicated time for addressing participant queries and debugging common issues.
- Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com