Understanding Apache Kafka (3 days course): Apache Kafka is one of the most used distributed data hubs in big data architectures. Its simple architecture and the way it writes and reads data makes it really fast at persisting and making available the billions of events. But what makes Kafka even more interesting is a number of services that can be deployed in conjunction with your Kafka cluster – Kafka Connect, KSQL, Schema Registry – services that transform Kafka from a distributed data bus into a distributed streams processing solution that makes it extremely easy to work with flexible schema formats like Apache Avro. In this course we will start from Kafka concepts and evolve through its architecture and then build on top of it adding components la Connect, Schema Registry and KSQL. While working with its ecosystem components we will as well lay the grounds of understanding what means processing live streams of data vs batch processing – we will work with persistent queries and understand what state means and how will be handled in distributed systems such as Kafka and KSQL.

This course is addressed to solutions architects, product managers or whomever would like to understand more about Apache Kafka and its ecosystem.

Main trainer: Crisan Valentina

Prerequisites students:

- For producing/consuming events we will mostly work with Kafka Console Producer/Consumer (KSQLdb client as well). While the course will address the Java API's for producers and consumers, this course is not addressed to devs thus anybody with an IT background can participate. **Knowledge of distributed systems/architectures is a plus.**

Structure of the course:

1. **Why Apache Kafka & use cases**
    - Positioning of Kafka in the Big Data architecture
    - ingestion layer overview
2. **Apache Kafka Fundamentals**
    - Topics
        - Key, Value, Partition, Offset, Timestamp & Timestamp Type
    - Producers: writing messages to Kafka
        - Overview, sending a message, configuring producers
        - Partitioning data
        - Hands on - producing data in topics
    - Consumers: reading from Kafka
        - Overview, subscribing to topics, offsets, configuring consumers
        - Consumer groups
        - Consumer group coordinator, group leader
        - Hands-on - consuming topics events
    - Brokers
        - Replication of data

        - Write / Read Path
        - Consistency vs Availability in Kafka
    - Data retention options
3. **Kafka Architecture**
        - Role of Zookeeper in Kafka clusters (depending on Kafka version)

4. **Kafka Connect**
        - Role of the Connect in Kafka
        - Architecture of a cluster with Kafka Connect
        - Connect Kafka to MySQL using Kafka Connect
        - Hands on - creating file systems source connectors and MySQL source/sink connectors
5. **Stream Processing**
        - Streams processing architectures concepts
            - Streams vs tables
            - Persistent queries
            - State handling in distributed systems
        - Kafka streams architecture overview
            - KSQL overview
            - KSQL concepts
            - Hands on using Kafka SQL:
                - create streams + tables
                - Operations with streams and tables: join stream & table
                - Handling different formats of data: JSON nested format, Avro using Schema Registry
            - Window aggregations
                - Handling of late data

Course Requirements:

- Each participant needs to have it's own computer in order to run the hands on exercises, also the computer settings has to allow **access to Google docs** and **Github** for getting access to presenters slides, documents, data and exercises;
- **Google Chrome browser;**
- **Each participant computer will need an SSH client to connect to the cloud environment.**
- All participants computers need **an open** Internet connection throughout the course, we will run the exercises on cloud - thus is mandatory for the Internet connection to be open and reliable;