DEEP DIVE INTO THE NEW FEATURES OF APACHE SPARK 3.0

MARIO CARTIA

Executive Managing Director, Agile Skill Entrepreneur

INTRODUCTION

Spark is a distributed general-purpose cluster-computing framework initially started by MateiZaharia at UC Berkeley's AMPLab in 2009 in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs. In 2010 it was released under the open source (BSD) license and in 2013 donated to the Apache Software Foundation. Today it is the "de facto standard" in Big Data Analytics thanks to its features that make it suitable for the most diverse areas ranging from ETL to Data warehouse up to the distributed training of artificial neural networks.
The latest version 3.0, released in June 2020, introduces a number of features aimed at increasing performance in various areas such as analytics (e.g. Adaptive Query Execution, Dynamic Partition Pruning), streaming (new UI for Structured Streaming) or deep learning (Accelerator-aware scheduling). The talk will introduce to the use of the framework, the most common use cases and the new features introduced with the latest version. 1 Keywords Big Data, Analytics, IoT, Artificial Intelligence, Machine Learning, Deep Learning.

KEYWORDS

Big Data, Analytics, IoT, Artificial Intelligence ,Machine Learning, Deep Learning

BIOGRAPHY

Mario Cartia Graduate in Information Security, Cloudera and Red Hat Certified Professional with +20 years experience in enterprise IT solutions and +10 inthe Big Data area. Consultant and trainer for some important international companies. Technical evangeli stand speaker at some of them ostimportant Italian technical conferences ,lecturer in schools and universities.