SPARK TRAINING COURSE

Paco Nathan

Apache Spark and the Emerging Technology Landscape for Big Data

This talk will present an overview about Apache Spark, along with its history and context within open source Big Data frameworks. Spark has become one of the most popular and active projects at the Apache Software Foundation. We will review a range of typical use cases and example how Spark puts data analytics into production at scale — including SQL queries federated from multiple sources, streaming applications, machine learning models, graph algorithms, etc.
We will then look toward innovations and impact on the industry in general based on Spark, considering where current research efforts are focused. Application areas such as genomics, IoT, and remote sensing have become drivers for innovation, compelling better methods for streaming algorithms, data visualization, fault-tolerant architectures, etc. As the IT industry grapples with an evolution of requirements for handling real-time data at scale from a variety of sources, how do Spark and complementary open source projects best fit into the emerging technology landscape?

Paco Nathan is the Director of Community Evangelism at Databricks, working on Apache Spark. He is also an O'Reilly author, and an advisor for the Amplify Partners investment portfolio and the GalvanizeU graduate program in data science.
With expertise in distributed systems, machine learning, functional programming, and cloud computing, Paco has led innovative Data teams building large-scale applications. He has 30+ years technology industry experience ranging from Bell Labs to early-stage start-ups, and was cited in 2015 as one of the Top 30 People in Big Data and Analytics by Innovation Enterprise. http://liber118.com/pxn/

David Martínez

Round table: Research in Big Data. Challenges

Big Data has changed the way we see the world and probably also the way we as researchers should see our role in society. This round table aims to be a review of the field of Big Data at a bird's eye view in order to disentangle what researchers are expected to bring to society in the years to come and what risks can prevent us from fulfilling that role.

David Martínez is a PhD in Computer Science specialised in designing tailored software systems and algorithms that include Machine Learning techniques in different areas like engineering, advertising, … In the past, he has been part of laboratories in different academic institutions such as University of A Coruña, University of Florida and University College London. Currently, he is based in London and shares his time between doing postdoc research and lecturing at UCL and tech consulting for different industries and start-ups. He is part of some scientific networks such as the Spanish National Network for Big Data and HPC and the Spanish Association of Artificial Intelligence. He is also a mentor for the UCL's Computational Finance apprenticeships programme. His main interests are Machine Learning theory and applications, software engineering and specially Big Data technologies and architectures.

Leslie Kanthan

Machine Learning in Finance

Introduction to FinTech, evolution of Technology, the interest in FinTech on ML and Big Data, Emerging RechTech ML, the Use cases at the commercial level, the future of ML and Big Data.

Leslie Kanthan is a consultant specialising in Quantitative Finance with 10 years of industry experience. He possess a background in Mathematics from the University of Warwick, an Advanced Masters with a major in Graph Theory from the London School of Economics and a Masters in Computational Statistics & Machine Learning from University College London. He is currently a PhD Research Student in Mathematics and Computer Science in the field of Graph Theory & Combinatorics with

SPARK TRAINING COURSE

focus on Partition Theory/Clustering complexity. Leslie began his career in Catastrophe forecasting. Later the use of Machine Learning began to enter the industry and he was one of the adopters of the use of ML in quantitative finance. Leslie's focus was preliminarily in the area of semi-supervised learning. He has consulted and worked for several tier 1 institutions: Microsoft Research, Orange Analytics, CommerzBank, Morgan Stanley, Credit Suisse, Santander Analytics, GMEX Group, Leehman Brothers and Bank of New York. In addition he has co-authored internal research papers at London Stock Exchange, GMEX Global Markets Exchange Group, Credit Suisse and Orange Analytics on Order book latency, derivative pricing, Prediction Analysis using Machine Learning in Big Data, and Queueing Theory respectively. He has also lectured Bond Mathematics, Financial Markets at UCL, run workshops in Game Theory at LSE, assisted in a documentary presentation of the mathematics of bitcoin for the BBC and has supervised over 60 individual masters students in Machine Learning, Mathematics and Computational Finance on projects/dissertation.

Juan Tomás García

All what wanted to know of Spark and did not dare to ask

Juan Tomás discovered his passion for technology when he was child. That passion has led him to explore new technologies and ended up hooked on BigData. Since then not stop talking about Apache Spark, Kafka, NoSQL, Architecture Lambda and / or Scala.
In ASPgems he is Big Data Manager and leader in Data Science and Data Management.
He leads ASPgems highly qualified technical teams, and design strategies and solutions for statistical analyses with success.
Co-author with Alfredo Romeoque of the book "La pastilla roja" that was the first book about free software in Spanish. He was president of Hispallinux and is an active member of the community of free software and open knowledge.

PROGRAMA

04/05/2015 11:00 Salón de Actos (Conferences room). Faculty of Informatics. UDC

- **11:00. Welcome:** Ricardo Cao Abad. Vice-chancellor of Investigation and Transfer. University of A Coruña.

- **11:30. Paco Nathan**: *Apache Spark and the Emerging Technology Landscape for Big Data.*

- **16:30. David Martínez:** *Round table: Research in Big Data. Challenges.*

- **18:00. Leslie Kantan:** *Machine Learning in Finance.*

SPARK TRAINING COURSE

PROGRAMA

05/05/2015 9:00-14:00
Faculty of Informatics. UDC

- **Juan Tomás García:** All what wanted to know of Spark and did not dare to ask.

- **09:00:** Introduction + Installation
- **09:20**: Spark Deconstructed
- **09:40**: A brief history
- **10:00**: Simple Spark Apps
- **10:20**: Spark Essentials
- **11:05**: Spark Examples
- **11:15**: (break)
- **11:30**: Unifying
- **12:15**: The full SDLC
- **13:30**: Case Studies + Follow Up

Registration / Requirements

Registration required by email:
- sparktraining15@gmail.com

Requirements:
- Laptop / Computer
- Software: Spark

Resources for the course are available at:
- databricks.com/spark-training-resources#itas

Download slides+code+data to your laptop:
- training.databricks.com/workshop/itas_workshop.pdf
- training.databricks.com/workshop/usb.zip

Organisation:

- Network of Technologies Cloud and Big Dates for HPC. (Coordinator Juan Touriño Domínguez)
- Spanish network of Big Dates and Analysis of scalable data (Coordinator: Francisco Herrera Triguelos)
- David Martínez Rego, UCL-UDC
- Amparo Alonso Betanzos, CITIC-UDC

XUNTA DE GALICIA
CONSELLERÍA DE CULTURA, EDUCACIÓN E ORDENACIÓN UNIVERSITARIA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL
UNA MANERA DE HACER EUROPA
Red BigDADE Big Data y Análisis de Datos Escalable
Rede de Tecnoloxías Cloud e Big Data para hpc
citic