PRESENTED BY O'REILLY AND CLOUDERA

MAKE DATA WORK
MARCH 13–14, 2017: TRAINING
MARCH 14–16, 2017: TUTORIALS & CONFERENCE
SAN JOSE, CA

Using R for scalable data analytics: From single machines to Hadoop Spark clusters

Join Attendee Network
Add to Your Schedule
Add Comment or Question

Vanja Paunic (Microsoft), Robert Horton (Microsoft), Hang Zhang (Microsoft), Srini Kumar (LevaData, Inc.), Mengyue Zhao (Microsoft), John-Mark Agosta (Microsoft), Mario Inchiosa (Microsoft), Debraj GuhaThakurta (Microsoft Corporation)

9:00am–12:30pm Tuesday, March 14, 2017

Data science & advanced analytics

Location: LL21 C/D

Level: Intermediate

Secondary topics: R

Average rating: ★★★☆☆ (2.50, 4 ratings)

Rate This Session

Download slides (PDF)

Who is this presentation for?

Data scientists, machine-learning scientists, and statisticians

Prerequisite knowledge

Programming experience in R

Familiarity with machine-learning algorithms

Materials or downloads needed in advance

A WiFi-enabled laptop with an SSH client with portforwarding capability (On MacOS or Linux, simply run the SSH command in a terminal window. On Windows, download and install plink.exe.)

What you'll learn

Learn how to perform scalable data science in R using appropriate compute infrastructure, distributed algorithms, out-of-memory computational techniques and access codes, and worked-out samples from public repositories and adopt them in practice

Description

R is one of the most used languages in the data science, statistics, and machine-learning (ML) community. Although open source R has a rich set of packages and functions for statistics and ML, when it comes to scalable data science, many CRAN-R users are hindered by the limitations of available functions to handle big data efficiently and a lack of knowledge about the appropriate computing environments to scale R scripts from single-node to elastic and distributed cloud services, including Spark 2.0 integrations.

Vanja Paunic, Robert Horton, Hang Zhang, Srini Kumar, Mengyue Zhao, John-Mark Agosta, Mario Inchiosa, and Debraj GuhaThakurta walk you through creating end-to-end data science solutions in R on Spark clusters and consuming them in production.

The tutorial materials and the scripts that are used to create the Spark clusters will be published to a public GitHub repository, so you'll be able to create Spark clusters identical to the ones you use in the tutorial by running the scripts even after the tutorial session completes.

Vanja Paunic

Microsoft

Vanja Paunić is a data scientist on the Azure Machine Learning team at Microsoft. Previously, Vanja worked as a research scientist in the field of bioinformatics, where she published on uncertainty in genetic data, genetic admixture, and prediction of genes. She holds a PhD in computer science with a focus on data mining from the University of Minnesota.

Robert Horton

Microsoft

Bob Horton is a senior data scientist in the Microsoft Partner ecosystem. Bob came to Microsoft from Revolution Analytics, where he was on the Professional Services team. Long before becoming a data scientist, he was a regular scientist (with a PhD in biomedical science and molecular biology from the Mayo Clinic). Some time after that, he got an MS in computer science from California State University, Sacramento. Bob currently holds an adjunct faculty appointment in health informatics at the University of San Francisco, where he gives occasional lectures and advises students on data analysis and simulation projects.

Website

Hang Zhang

Microsoft

Hang Zhang is a senior data science manager on the Algorithm and Data Science team in the Data group at Microsoft, where his major focus is on team data science processes and the Cortana Intelligence Competition Platform. Previously, Hang was a staff data scientist at WalmartLabs in charge of internal business intelligence tools and a senior data scientist at Opera Solutions. He is a senior member of the IEEE. Hang holds a PhD in industrial and systems engineering and an MS in statistics from Rutgers University.

Srini Kumar

LevaData, Inc.

Srini Kumar is the vice president of product management and data science at LevaData, Inc. Previously, he was a director of data science in the Algorithms and Data Science group at Microsoft, where he worked with strategic customers in the areas of Cortana Analytics and Microsoft R Server; headed product management for the information management (EIM) product suite at SAP; originated and architected a product on HANA to analyze human genome variants, which led to a discovery relating diabetes to a person's origin and resulted in two patent applications related to modeling genomic variants and one related to enterprise information management; and helped turn around and sell a startup in the area of on-demand supply chain management software. Srini holds a master's degree in industrial engineering from the University of Wisconsin-Madison and a bachelor's degree in mechanical engineering from the Indian Institute of Technology, Madras.

Mengyue Zhao

Microsoft

Mengyue Zhao is a data scientist at Microsoft, where she develops end-to-end machine-learning solutions for various use cases in cloud computing and distributed platforms (e.g., Azure, Hadoop, and Spark). Mengyue focuses on scalable analysis, including data processing, feature engineering, feature selection, predictive modeling, and web services development. Previously, she was a data analyst at GE Digital, mainly focusing on solving machine-learning problems in the manufacturing domain. Mengyue has broad interests in machine learning, deep learning, and data mining and is passionate about harnessing the power of big data to answer interesting questions and drive business decisions. Mengyue holds a master's degree in analytics from the University of San Francisco.

John-Mark Agosta

Microsoft

John Mark Agosta is a principal data scientist in IMML at Microsoft. Over his career, he has worked with startups and labs in the Bay Area, including the original Knowledge Industries, and was a researcher at Intel Labs, where he was awarded a Santa Fe Institute Business Fellowship in 2007, and at SRI International after receiving his PhD from Stanford. He has participated in the annual Uncertainty in AI conference since its inception in 1985, proving his dedication to probability and its applications. When feeling low he recharges his spirits by singing Russian music with Slavyanka, the Bay Area's Slavic music chorus.

Mario Inchiosa

Microsoft

Mario Inchiosa's passion for data science and high-performance computing drives his work at Microsoft, where he focuses on delivering parallelized, scalable advanced analytics integrated with the R language. Previously, Mario served as

Revolution Analytics's chief scientist and as analytics architect in IBM's Big Data organization, where he worked on advanced analytics in Hadoop, Teradata, and R. Prior to that, Mario was US chief scientist in Netezza Labs, bringing advanced analytics and R integration to Netezza's SQL-based data warehouse appliances. He also served as US chief science officer at NuTech Solutions, a computer science consultancy specializing in simulation, optimization, and data mining, and senior scientist at BiosGroup, a complexity science spin-off of the Santa Fe Institute. Mario holds bachelor's, master's, and PhD degrees in physics from Harvard University. He has been awarded four patents and has published over 30 research papers, earning Publication of the Year and Open Literature Publication Excellence awards.

Debraj GuhaThakurta

Microsoft Corporation

Debraj GuhaThakurta is a senior data scientist in Microsoft's Azure Machine Learning group, where he focuses on the use of different platforms and toolkits, such as Microsoft's Cortana Analytics Suite, R Server, SQL Server, Hadoop, and Spark clusters, for creating scalable and operationalized analytical processes for various business problems. Debraj has extensive industry experience in the biopharma and financial forecasting domains. He holds a PhD in chemistry and biophysics and did postdoctoral research in machine-learning applications in genomics. Debraj has published more than 25 peer-reviewed papers, book chapters, and patents.

Website

Leave a Comment or Question

Help us make this conference the best it can be *for you*. Have questions you'd like this speaker to address? Suggestions for issues that deserve extra attention? Feedback that you'd like to share with the speaker and other attendees?

Join the conversation here (requires login)

Presented by

cloudera O'REILLY

Elite Sponsors

MAPR Microsoft

Strategic Sponsors

Google Cloud IBM intel MEMSQL

TERADATA

Zettabyte Sponsor

DATASCIENCE

Contributing Sponsors

bmc

DELLEMC

informatica

Paxata

pentaho A Hitachi Group Company

SAP

ZALONI THE DATA LAKE COMPANY

Exabyte Sponsors

amazon web services

DataRobot

HUAWEI

HORTONWORKS POWERING THE FUTURE OF DATA

RYFT

sas

snowflake

talend

TalkingData vmware VOLTDB

Terabyte Sponsors

ClearStory DATA Now You See It GitHub

Impact Sponsors

confluent deepsense.io BIG DATA SCIENCE IMPETUS kinetica

kyvos insights pepperdata SECURONIX Security. Analytics. Delivered. striim

syncsort TRILLIUM SOFTWARE TRIFACTA

Supporting Sponsor

wanDISCO

Community Partners

The Apache Software Foundation http://www.apache.org/ CIOReview Cool Infographics CTOvision.com

Sponsorship Opportunities

For exhibition and sponsorship opportunities, email stratahadoop@oreilly.com

Partner Opportunities

For information on trade opportunities with O'Reilly conferences, email partners@oreilly.com

Contact Us

View a complete list of Strata + Hadoop World contacts

Information
About
Diversity
Code of Conduct
Privacy Policy
Contact Us

More O'Reilly Events
Artificial Intelligence
Design
Fluent
JupyterCon
Next:Economy
Open Source
Security
Software Architecture
Velocity

More O'Reilly Sites
Safari
O'Reilly Conferences
oreilly.com
O'Reilly Video Training
O'Reilly Webcasts
O'Reilly on YouTube

Twitter
Google+
Facebook
LinkedIn
YouTube