From Big Data to Social Media Analytics

Last Updated Friday, 19 June 2015

SyllabusIntroduction (1h)

Content
- Why now?

- What is Big Data? volume, velocity, variety, veracity, … , and value
- Paradigm shifts enabled
- Market Landscape

Material
- a slightly revised version of the slides [pdf] I preseted at the Politecnico di Milano PhD course on A Multidisciplinary Perspective On Big Data

Mastering the Volume Dimension (9h)
Operational perspective: the NoSQL world (3h)

Content
- Recalling basic concepts of the relational model
- NoSQL: basic concepts
- Key-Value stores

- Column-family stores
- Hands-on HBase
- Document-based stores

- Hands-on MongoDB (?)

- Graph-based stores

Material
- the slides [pdf] that prof. E. Di Nitto preseted at the Politecnico di Milano PhD course on A Multidisciplinary Perspective On Big Data
- HBase
- download from an italian mirror site

- getting started using the official guide

- MongoDB
- download from the official site

- getting started using the official guide

Analytical perspective: from Map Reduce (hadoop) ... (2h)Content

- fundamentals
- pros and cons
- evolution

- eco-system

Material
- the slides [pdf] that prof. D. Ardagna preseted at the Politecnico di Milano PhD course on A Multidisciplinary Perspective On Big Data Analytical perspective: ... to Spark (4h)Content

- Introduction
- Hands-on: simple Apps

- Essentials
- Hands-on Spark SQL
- Hands-on MLlib, k-means
- Hands-on GraphX

Material
- a selection of the slides and software packages (2.1GB!) in the Spark Summit 2014 Training Archive

Mastering the Variety Dimension (1h30m)Content

- Variety is unavoidable
- Embrace variety with semantic technologies
- Demonstration of ontop

Material
- my own slides on mastering the Variety dimension of Big Data [pdf]
- demonstration of ontop

- my own slides demonstrating of ontop [pdf] adapted from a practice session of Politecnico di Milano course on ICT for Healthcase
- the ontologies, mappings, queries and relational data [zip] that I prepared for the demonstration

- download Protege 5.0 + -ontopPro-: Protege 5.0 bundled with -ontopPro- and the JDBC plugins. This is ready to run package, use the run.sh or run.bat start scripts.
- download H2 + ontop tutorial databases: H2 database server bundled with the databases used in the ontop tutorials. Use this to avoid having to install a database to run the tutorials.Mastering the Velocity Dimension and beyond (2h30m)Content

- It's a streaming world
- Information flow processing
- Hands-on Event Processing Language
- Volume+Velocity: Hands-on Spark Streaming
- Velocity+Variety: Stream Reasoning

Material
- a slightly revised version of the slides [pdf] I preseted at the Politecnico di Milano PhD course on A Multidisciplinary Perspective On Big Data
- an EPL-centric version of the slides [pdf] I will present at the Politecnico di Milano PhD course on Complex Event and Stream Processing
- an online tool to tryout EPL on esper [link]
- the stream reasoning Web site [link] The "Traditional" volume-centric use cases and case studies (2 hours)

Content
- The Forrester WaveTM to choose among the vendors
- Amazon
- Cloudera
- Hortonworks
- MapR technologies
- IBM Big Insights
- Microsoft Azure HDInsight

Material
- my brand new slides (value the links in the presentation!) Social Media Analytics (5 hours)

Content
- what's social media?
- why shuould I care?
- Personal Social Media Analytics
- Social Media Analytics for companies
- So, where's Big Data?

- From micropost to (Big) data
- Collecting Social Media

- Named Entity Recognition
- Entity Linking
- Sentiment/Opinion extraction
- Example of Social Media Analytics

Material
- my brand new slides (value the links in the presentation!)

- Example of Social Media Analytics from the Stream Reasoning group of Politecnico di Milano and Fluxedo The 5

game changing big data use cases (6 hours)

Content
- Introduction to the 2015 study by IBM Analytics on the five high-value use cases that can be the first step into big data
- the study
- how to read it
- Use cases solved with IBM technology as well as with other products offered on the big data market
- Big data exploration use case
- Enhanced 360-degree view of the customer use case
- Security/intelligence extension use case
- Operations analysis use case
- Data warehouse modernization use case
- Conclusions

Material
- my brand new slides (value the links in the presentation!)City Sensing case study (2 hours)

Content:
- The digital reflection of our cities is sharpening and it is tracking their evolution with a decreasing delay. This happens thanks to the pervasive deployment of sensors, the wide adoption of smart phones, the usage of (location-based) social networks and the availability of datasets about urban environment.

- So while data becomes every day more abundant, decision makers face the challenge to increase their capability to create value out of the analysis of this data.

- This part of the course presents how advance visual analytics, ontology base data access and information flow processing methods can help in making sense of Social Media Streams and Call Data Records from Mobile Network Operators during city scale events. Real-world deployments demonstrate the ability of those methods to advance our ability to feel the pulse of our cities in order to deliver innovative services.

Material
- the key note I will give at BIS 2015
Spark case studies (1 hour)

Content
- Spark at Twitter
- Hadoop and Spark Join Forces inYahoo
- Collaborative Filtering with Spark at Spotify
- Stratio Streaming: a new approach to (Spark Streaming!
- Sharethrough Uses Spark Streaming to Optimize Bidding in Real Time
- Guavus Embeds Apache Spark (into its Operational Intelligence Platform (Deployed at the World’s Largest Telcos
- One platform for all at Conviva: real-time, near-real-time, (and offline video analytics on Spark
- others from Spark summit 2015

Material
- a subset ot the slides in in the Spark Summit 2014 Training Archive