

Distributed Enabling Platforms (PAD)



- Teacher(s) name: Nicola Tonellotto

email: nicola.tonellotto@isti.cnr.it

tel: 050 315 2967

web: <http://hpc.isti.cnr.it/~khast/>

- Semester: 1st
- Exam mode: project + oral examination

Infrastructure

Hadoop On-Premise
cloudera Hortonworks MAPR Pivotal IBM InfoSphere bluedata Jethro

Hadoop in the Cloud
amazon Microsoft Azure Google Cloud Platform IBM InfoSphere CAZENA altiscale

Spark
databricks GridGain TACHYON NEXUS

Cluster Services
amazon web services Substrata MESOSPHERE CoreOS Docker StackIQ

Analytics

Analyst Platforms
Palantir AYASDI Quid enigma Digital Reasoning ORBITALINSIGHTS

Analytics Platforms
Microsoft guavus Datameer Bottlenose interana

Data Science Platforms
context relevant CONTINUUM DataRobot Alpine MODE plotly dataiku Ionian DOMINO sense yhat ALGORITHMIA

Visualization
tableau Qlik looker Roambi Sisense QlikView Chatterbox Chatterbox CHARTIO

Applications

Sales & Marketing
RADIUS Gainsight bloomreach Zeta EVERSTRING blueyonder Lattice kahuna infer SAILTHRU persado AVISO sense QUANTIFINDO ACTIONIQ fuse:machines ENGAGIO

Customer Service
MEDALLIA ATTENTIVITY CLARABRIDGE CLICKFOX STELLASERVICE NG@DATA Preact DigitalGenius appurion Wiseio

Human Capital
gild Connectifier textIQ entelo hiQ

Legal
RAVEL JUDICATA Everlaw Brevia PROMOBATION

NoSQL Databases

amazon DynamoDB Google Cloud Platform Microsoft Azure ORACLE mongoDB MarkLogic DATASTAX Couchbase KERO SPIKE SequoiaDB redislabs Influxdata

NewSQL Databases

SAP Clustrix Pivotal paradigm4 nuodb memsql VOLTDB splicewise MariaDB citusdata deepdb Trafodion Cockroach LABS

BI Platforms

Power BI amazon web services Domo Wave Analytics GoodData birst platform atscale

Statistical Computing

sas SPSS MATLAB

Log Analytics

splunk sumologic kibana CLOUD PHYSICS loggly

Social Analytics

Hootsuite NETBASE DATASIFT track bitly synthesio simplereach

Graph Databases

neo4j OrientDB InfoGraphs

MPP Databases

TERADATA VERTICA NETEZZA COTION Kognitio XSQL Greenplum

Cloud EDW

amazon web services Google Cloud Platform Microsoft Azure Pivotal snowflake MATRIUM DATA Infoworks

Data Transformation

alteryx talend TRIFACTA tamr StreamSets Alation

Data Integration

informatica Full potential to work MuleSoft snapLogic BedrockData xplenty

Real-Time

amazon web services METAMARKETS striim confluent DATATONHERBY dataArtisans

Machine Learning

Azure ML Learning H2O Dato SKY TREE rapidminer DATAFLOW deep2zero VISERZ Predictator.io glowfish

Speech & NLP

NarrativeScience NUANCE semantic processor ARRIA apical MindMeld IDIBON VSCOOP

Horizontal AI

IBM Watson Cortana sentient VIV nano Numenta clarifai MetaMind

Management / Monitoring

New Relic APPDYNAMICS amazon web services actriov Numerity splunk DATA DOGS DRIVEN Anodot

Security

TANIUM illumio CODE42 DataGravity CipherCloud VECTRA sqrrl BlueTalon

Storage

amazon web services Google Cloud Platform Microsoft Azure panasas nimblestorage COHO Qumulo

App Dev

apigee CASK Typesafe DRIVEN

Crowd-sourcing

amazon mechanicalturk CrowdPower WorkFusion

Search

hp Oracle ENDeca EXALEAD Lucidworks elastic ThoughtSpot MAANA swifttype Algolia SHREQUA

Data Services

UC OPERA Mu Sigma EXL KAGGLE data science kaggle dataKIND

For Business Analysts

OrigamiLogic ClearStory CIRRO Import.io

Web / Mobile / Commerce

Google Analytics mixpanel RjMetrics BLUECOSE AMPLITUDE granify sumal Airtable retention custora

Publisher Tools

Outbrain Taboola quantcast Chartbeat yieldbot Yieldmo

Govt / Regulation

Socrata OPENGOV EN FiscalNote PREDPOL enigma mark43 OpenDataSoft

Finance

affirm LendingClub OnDeck Kreditech Kabbage tidemark INSIKT UORO Dataminr Lenddo KENSHO AIDYIA iSENTIUM Quantopian sentient

Education / Learning

KNEWTON Clever Cleclara PANORAMA knowre

Life Sciences

23andMe Counsyl Recombinome KYRUS FLATIRON oozymyrgen HealthTop METABIOTA ZEPHYR HEALTH OVIQ Gingerio transcriptic Glow enitic AiCure Atomwise

Industries

OP@WER eHarmony RetailNext duetto STITCH FIX BLUE@RIVER WorkFusion TACHYUS SwiftKey SeeQ FarmLogs HowGood select RIGHT MACHINES statmuse B@XEVER

Cross-Infrastructure/Analytics

amazon web services Google Microsoft IBM SAP SAS SAS ITO data hp VMware VERTICA vmware TIBCO TERADATA ORACLE NetApp

Open Source

Framework: Hadoop HADOOP HADOOP YARN Spark MESOS TEZ Flink CDAP

Query / Data Flow: SLAMDATA HIVE Google Cloud Dataflow

Data Access: cassandra HBASE mongoDB CouchDB riak OPENSTACK nifi

Coordination: talend Apache Zookeeper Apache Ambari

Real-Time: STORM Spark APEX Flink TAGYON druid

Stat Tools: ScalaLab SciPy

Machine Learning: mllib Apache SINGA MADlib Aerolve Caffe FeatureFu DIMSUM VELES WEKA jupyter DL4J

Search: elasticsearch Solr

Security: Apache Ranger

Visualization: Kognitio

Data Sources & APIs

Health: Apple JAWBONE GARMIN practicefusion fitbit Withings VALIDIC nestatmo kinso Human API

IOT: UPTAKE ThingWorx beium samsara

Financial & Economic Data: Bloomberg DOW JONES THOMSON REUTERS YODLEE PREMISE S&P CAPITAL IQ quandl xignite CBINSIGHTS mattermark estimize PLAID

Air / Space / Sea: PLANET LABS spire WINDWARD CRUISE SKY CATCH Airware DroneDeploy

Location / People / Entities: acxiom Experian EPSILON GARMIN foursquare InsideView esri STREETLINE CAIROOB factual PlaceIQ Climbin Hexagon placemeter BASIS Senso

Other: qualtrics panjiva DATA.GOV

Incubators & Schools: GA PLURAL SIGHT DataCamp INSIGHT DataElite The Data Incubator METIS



Syllabus



- Design issues and solutions in very-large-scale distributed systems
- The objectives of this course are:
 - to develop an understanding of the typical issues of very large scale distributed systems;
 - to equip students with tools, best practices and common procedures to design, implement and program such systems, through understanding of algorithms and suitable theoretical models.
- List of topics:
 1. Introduction to large scale distributed systems.
 2. Cloud computing: introduction, service and deployment models, solutions.
 3. Infrastructure: virtualization, coordination, scalability, availability
 4. Programming: mapreduce model, APIs, patterns
 5. Data: data management, consistency, replication, fault tolerance.

Expected Outcomes



- Distributed Computer System Engineer
 - Analyze requirements
 - Understand design choices
 - Propose and implement solutions
- Large-scale Data Manager
 - Data warehousing
 - Analysis of solutions
- Big Data Analyst
 - Program applications to crunch terabytes of data
 - From numbers to texts to structured data

Thesis available



- Web search algorithms for efficient processing
- Distributed and replicated architectures for Web search
- Energy-efficient Web Search
- Green Information Retrieval
- Dynamic modeling of distributed systems
- Large-scale algorithms for data processing
- Efficient large-scale machine learning algorithms