CLOUD COMPUTING AND BIG DATA
INF/01 - 6 CFU - 2° semestre

Docente titolare dell’insegnamento
GIUSEPPE PAPPALARDO
Email: pappalardo@dmi.unict.it
Edificio / Indirizzo: Dipartimento di Matematica e Informatica, Blocco 3, Ufficio n. 48, Viale A. Doria
6, Catania
Telefono: 0957383000
Orario ricevimento: http://www.dmi.unict.it/pappalardo/ricevimento.php

OBIETTIVI FORMATIVI

1. **Knowledge and understanding (Conoscenza e capacità di comprensione).** Students will acquire a precise knowledge and understanding of fundamental concepts in the field of cloud computing, chiefly through a guided exploration of the main technological solutions available from the public Cloud, focusing on resources and services oriented to data storage, analysis, visualization and machine learning.

2. **Applying knowledge and understanding (Capacità di applicare conoscenza e comprensione).** Based on the operating knowledge acquired, students will develop an effective "toolset" of practical, application-oriented skills in leveraging the Cloud to cater for the typical needs of a data scientist: i.e. processing large datasets with a view to revealing meaningful patterns and relationships. Cloud implementations of state-of-the-art tools and frameworks like, e.g., MapReduce/Hadoop or TensorFlow, will be employed.

3. **Making judgements (Autonomia di giudizio).** The student will develop the ability to choose the suitable Cloud-based resource for the Data Science scenario of interest, properly estimating the ensuing costs and performance gains, as well as consciously assessing the tradeoffs involved.

4. **Communication skills (Abilità comunicative).** The student will acquire the communication skills required to express and discuss, at a rigorous technical level, the benefits and (mostly cost-related) downsides of the Cloud for Data Science applications. In addition, the student will gain the ability, for presentation purposes, to effectively highlight the features of very large datasets by means of cloud-based visualization services.

5. **Learning skills (Capacità di apprendimento).** Students will become capable of profitably consulting technical documentation concerning Data Science-oriented Cloud services, in order to concretely put them to effective use.

MODALITÀ DI SVOLGIMENTO DELL’INSEGNAMENTO

Lectures will mainly consist in live sessions dealing with using the Cloud for the purposes of data analysis.
and machine learning. These sessions will be carried out by the lecturer and replicated, with suggested variations, by the students, on available equipment. Laboratory practice aims at enabling students to refine their understanding of the technologies presented and acquire autonomous operating skills. As a framework and guidance, lecture notes will be displayed during lectures and shared with students. Notes will provide a precise record of the material presented, as well as pointers to the required reference technical documentation.

Should teaching be carried out in mixed mode or remotely, it may be necessary to introduce changes with respect to previous statements, in line with the programme planned and outlined in the syllabus. Learning assessment may also be carried out on line, should the conditions require it.

**PREREQUISITI RICHIESTI**


**FREQUENZA LEZIONI**

Attending classes is not mandatory but strongly recommended.

**CONTENUTI DEL CORSO**

This course aims at enabling the data scientist to put into practice on the public Cloud principles and methodologies learnt in courses concerned with data storage, processing, analysis, and machine learning. Indeed, in these areas, present day industrial and enterprise applications typically require storage volumes, computing power and bandwidth at a scale impossible or (even for large organizations) impractical to attain with proprietary equipment. In realistic Data Science scenarios, it is therefore hardly avoidable for the data scientist to resort to the Cloud, i.e. storage and computing services offered by third-party providers over the public Internet, with a pay-per-use cost model.

In a nutshell, quoting reference [2], we may say that: “The Cloud turbocharges Data Science”.

Google Cloud is the platform of choice, for its ease of use and free availability to students.

**SQL on Google Cloud and BigQuery:** performing structured queries on BigQuery and Cloud SQL. Importing data from CSV files.

**Data acquisition into Google Cloud:** downloading selected data from a large public data set over the internet, and processing it with Google App Engine.

**Google Cloud Dataflow:** processing a real-time, real-world data set, and storing the results on the cloud. Case study: real-time geospatial data.

**Visualization with Google Data Studio:** Visualizing data stored in Google Cloud SQL. Visualizing Real Time Geospatial Data.

**Google Datalab for Data Analysis:** loading text data into Google BigQuery; rapid exploratory data analysis with Google Cloud Datalab notebooks.
Google Cloud AI Platform: using Google AI Platform to perform queries and present the data.

Evaluating a Data Model: partitioning a data set into a training set and a test set; evaluating various predictive models.

Machine Learning with Spark on Google Cloud Dataproc. Implementing logistic regression through machine learning on Apache Spark running on a Google Cloud Dataproc. Developing a model from a multivariable dataset.

Machine Learning with TensorFlow: developing and evaluating prediction models.

MapReduce e Hadoop on Google Cloud: exploiting parallelism and machine clusters.

TESTI DI RIFERIMENTO

3. Lecture notes, to be made available through the Studium portal.

ALTRO MATERIALE DIDATTICO

Lecture notes will be made available through the Studium portal.

PROGRAMMAZIONE DEL CORSO

<table>
<thead>
<tr>
<th>Argomenti</th>
<th>Riferimenti testi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Google Cloud (GC): Performing structured queries on BigQuery</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>2 GC: Performing structured queries on Cloud SQL</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>3 GC: Importing big data from CSV files</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>4 Downloading large public data sets to GC</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>5 Processing data with the Google App Engine</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>6 GC Dataflow: processing a real-time, real-world data set</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>7 Case study: real-time geospatial data on GC</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>8 GC Data Studio: Visualizing data from Google Cloud SQL</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>9 GC Datalab: Data Analysis and Google BigQuery</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>10 GC Datalab notebooks for rapid exploratory data analysis</td>
<td>Lecture notes</td>
</tr>
<tr>
<td>11 GC AI Platform: queries and data presentation</td>
<td>Lecture notes</td>
</tr>
</tbody>
</table>
Partitioning a data set into a training set and a test set

Predictive models and their evaluation

Machine Learning (ML) with Spark on GC

ML with Spark on GC: implementing logistic regression

ML with Spark on GC: Developing a model from a multivariable dataset

ML with TensorFlow on GC: developing and evaluating predictive models

ML with TensorFlow on GC: developing and evaluating predictive models

MapReduce e Hadoop on Google Cloud: exploiting parallelism and machine clusters

VERIFICA DELL'APPRENDIMENTO

MODALITÀ DI VERIFICA DELL'APPRENDIMENTO
Laboratory session individually performed by the student vis-à-vis the lecturer. The student will be required to carry out the Cloud-based procedures demonstrated during the lectures, as well as to discuss their significance, and critically assess their outcomes. Learning assessment may also be carried out online, should the conditions require it.

ESEMPI DI DOMANDE E/O ESERCIZI FREQUENTI
See material available on the Studium portal.