

A screenshot of a Microsoft Teams meeting interface. The main content area displays a presentation slide with the following text:

Corso di Sistemi e Architetture per Big Data
A.A. 2019/2020
Valeria Cardellini, Fabiana Rossi

The slide has a dark background with white text. Below the slide, the Teams interface shows the following controls:

< Diapositiva 1 di 11 > Assumi il controllo ⚡ Torna al relatore ⚡

Indietro Valeria Cardellini

On the right side of the slide, there is a list of participants with their names and small profile pictures. At the bottom right of the slide, there is a decorative image of a rocky landscape. The overall interface includes various Teams navigation icons like Chat, Team, Activities, Calendar, Calls, File, and more.

The screenshot shows a Microsoft Teams meeting interface. The main area displays a presentation slide with the title "Project delivery". The slide content includes:

- Submission deadline
 - May 29, 2020
 - After the deadline, the maximum achievable score will be decreased by 2 points for each week of delay
- Your presentation
 - June 4, 2020
- What to deliver
 - Link to cloud storage or repository containing project code
 - Project report composed by 3-6 pages in ACM or IEEE proceedings format (only the report: by June 1)
 - Presentation slides (max. **15 minutes** per group), to be delivered after your presentation
- Team
 - 2 students per team
 - Also possible 1 student

At the bottom of the slide, it says "V. Cardellini, F. Rossi - SABD 2019/2020".

On the right side of the screen, there is a "Chat della riunione" (Meeting Chat) window. It shows a message from "Riunione" stating "Registrazione iniziata" (Recording started). Below this, there is a list of participants with their names and profile pictures. At the top of the chat window, there are buttons for "Informativa sulla privacy" (Privacy information), "Ignora" (Ignore), and "Rispondi" (Reply).

At the very bottom of the interface, there are navigation buttons for the presentation slide, including arrows for previous/next, a search bar, and other controls.

The screenshot shows a Microsoft Teams meeting interface. The main window displays a presentation slide titled "Dataset" with the following bullet points:

- You will use two real datasets on Covid-19
- Both datasets are available in CSV format
- Datasets are available at
 - <https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv>
 - https://csse.covid19data.piomonte.it/time_series/time_series_covid19_confirmed_global.csv

The Teams interface includes a top bar with the date (Lun 4 mag 12:29:31), a search bar, and various icons. A sidebar on the left shows team members and activity. A "Chat della riunione" (Meeting Chat) window is open, showing a message from "Francesco Marino" with the text "Informativa sulla privacy" (Privacy information) and a "Riunione Registrazione iniziata" (Meeting Recording started). The bottom right corner shows a video call interface with participant icons and a "VC" button.

The screenshot shows a Microsoft Teams meeting interface. The top bar includes icons for file, back, forward, and search, along with the title "Microsoft Teams" and the date "Lun 4 mag 12:29:34". A message in the center says "Esegui una ricerca o digita un comando". On the right, there's a "Chat della riunione" section with a purple button "Informativa sulla privacy" and a "Riunione" button with the text "Registrazione iniziata". The main content area has a dark background and displays the following slide:

Dataset 1: schema

- Input in CSV format
- One row per day, starting from February 24, 2020
- On each row:
 - Date with format yyyy-mm-ddThh:mm:ss
 - Country (ITA)
 - Patients with symptoms
 - Patients in intensive care units
 - Hospitalized patients
 - Home isolation
 - Total number of positives
 - Change of total confirmed cases on a daily basis
 - Number of new confirmed cases
 - Number of hospitalized and discharged (cured)
 - Total number of deaths

On the right side of the slide, the text "Most data are cumulative" is displayed.

At the bottom of the slide, there are navigation buttons: "Diapositiva 4 di 11", "Assumi il controllo", and "Torna al relatore". There is also a red box containing the number "6" and the text "Number of confirmed cases".

The bottom right corner of the slide shows the name "V. Cardellini, F. Rossi - SABD 2019/2020".

The bottom of the screen shows the Microsoft Teams ribbon with tabs for "Azione", "Team", "Calendario", "Attività", "Chat", and "File". The "Team" tab is selected. The status bar at the bottom right shows "Valeria cardellini", "App Guardia", "Guida", and "Diapositiva 3 di 11".

The screenshot shows a Microsoft Teams meeting interface. At the top, there's a navigation bar with icons for Microsoft Teams, Modifica, Visualizza, Finestra, Actions, Chat, Team, Activities, Calendario, Chiamate, File, and three dots. A status bar indicates it's Monday, May 4, 2020, at 12:29:41, and the user is Francesco Marino.

A registration pop-up window titled "Chat della riunione" is open, showing "Riunione" and "Registrazione iniziata".

The main content area displays a presentation slide with the title "Dataset 2: schema". The slide contains the following bullet points:

- Input in CSV format
- One row per day, starting from January 22, 2020
- On each row:
 - Province/State
 - Country
 - Latitude and longitude
 - Daily columns with total number of confirmed cases

At the bottom of the slide, there are navigation controls: "Diapositiva 5 di 11", "Assumi il controllo", and "Torna al relatore".

To the right of the slide, a participant list is visible, showing users like Valeria Cardellini, Marco Ballotti, Ilenia Rocca, and others, along with their profile pictures and names. There are also various application icons and a "Guarda" button.

The screenshot shows a Microsoft Teams meeting interface. The top bar displays the title 'Chat della riunione' and the date 'Lun 4 mag 12:29:49'. A message from Federica says 'La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.' Below the message, there's a button 'Informativa sulla privacy'.

The main content area features a slide with the title 'Queries with Hadoop/Spark'. The slide contains the following text and bullet points:

Esegui una ricerca o digita un comando

Δ La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.

Queries with Hadoop/Spark

- Use the Hadoop framework (and the MapReduce programming model) or alternatively the Spark framework to answer some queries on the dataset
- Include in your report/slides the queries' response time on your reference architecture

The right side of the screen shows the team members in the meeting, their status (e.g., 'VC', 'Marco Ballestri'), and various application icons for communication and collaboration.

The screenshot shows a Microsoft Teams meeting interface. At the top, there's a navigation bar with icons for Microsoft Teams, Modifica, Visualizza, Finestra, Actions, Chat, Team, Activities, Chiamate, Calendario, File, and three dots. The title bar indicates it's a meeting with Francesco Marino, dated Lun 4 mag 12:35:24. A message says "Esegui una ricerca o digita un comando". A participant message says "La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione." A button to "Informativa sulla privacy" is present.

The main area displays a presentation slide titled "Queries with Hadoop/Spark". The slide content is as follows:

Query 2

Using dataset 2, for each continent, determine the average, standard deviation, minimum and maximum number of confirmed cases on a weekly basis

- Consider only states with at least 50 confirmed (if State is not indicated, use Country)
- Use **trendline coefficient**
- Continent must be identified

At the bottom of the slide, there are navigation controls: Diapositiva 7 di 11, Assumi il controllo, Torna al relatore 7, and a back arrow.

On the right side of the screen, there's a "Chat della riunione" window with a message from Valeria Cardellini: "Rispondi", "Copia", "Stampa", "Invia", "...". Below the chat is a list of participants: Valeria Cardellini (VC), marco ballotti, ienia rocca, giacomo gastro, and Valeria Cardellini again. To the right of the participants are various application icons: App, +6, AC, Guarda, and a magnifying glass icon over a document. The background of the Teams interface features a wooden texture.

The screenshot shows a Microsoft Teams meeting interface. On the left, the Teams navigation bar is visible with icons for Actions, Chat, Team, Activities, Calendario, Chiamate, File, and three dots. The main area displays a presentation slide with the title 'Queries with Hadoop/Spark'. Below the title, the text reads: 'Using dataset 2 and the top-50 states identified using the trendline coefficient, use K-means clustering algorithm (with K=5) to identify the states and nations that belong to each cluster' followed by a bullet point: '– Compare the performance of a naïve implementation of K-means with that provided by Spark MLlib or Apache Mahout'. To the right of the slide, a participant's view of the meeting is shown. The participant's name is Valeria Cardellini, indicated by a blue circular icon with 'VC'. The participant's video feed shows a person with short hair. The participant's status is 'Rispondi' (Reply). The bottom of the participant's view shows a toolbar with icons for App, Guarda (View), +6, AC, and three dots. The background of the participant's view features a textured wooden surface.

The screenshot shows a Microsoft Teams interface. At the top, there's a navigation bar with icons for Actioni, Chat, Team, Attività, Calendario, Chiamate, File, and three dots. Below the navigation bar, there's a search bar and a message "Esegui una ricerca o digita un comando". A notification bar says "La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione." On the left, there's a sidebar with "Diapositiva 9 di 11" and "Assumi il controllo". In the center, there's a large dark area with the title "Optional part" in white. To the right, there's a "Chat della riunione" window with a purple header "Riunione" and a blue header "Informativa sulla privacy". The bottom right corner shows a list of icons for various applications like App, Guarda, and others.

Optional part

- Compulsory for team composed of 3 students
- Use either Hive (or Pig) or Spark SQL to address the Queries 1 and 2
- Include in the report the query times using a higher level framework on your reference architecture and compare them to those achieved by your pure Hadoop/Spark-based solution

The slide has a dark background with a wooden texture on the right side. It features a title "V. Cardellini, F. Rossi - SABD 2019/2020" and a footer "Valeria Cardellini". There are several circular icons with initials: "AC", "+6", "LC", "giacomo gastro", "ilena rocca", "marco ballotti", and "Valeria Cardellini". At the bottom, there are icons for "App", "Guarda", and three dots.

Diapositiva 9 di 11 > Assumi il controllo ⏪ Torna al relatore 10

Valeria Cardellini

App Guarda ...

+6 AC LC giacomo gastro ilena rocca marco ballotti Valeria Cardellini

Diapositiva 9 di 11 > Assumi il controllo ⏪ Torna al relatore 10

Valeria Cardellini

App Guarda ...

+6 AC LC giacomo gastro ilena rocca marco ballotti Valeria Cardellini

Diapositiva 9 di 11 > Assumi il controllo ⏪ Torna al relatore 10

Valeria Cardellini

App Guarda ...

+6 AC LC giacomo gastro ilena rocca marco ballotti Valeria Cardellini

The screenshot shows the Microsoft Teams application interface. At the top, there's a navigation bar with icons for Microsoft Teams, Modifica, Visualizza, Finestra, Actions, Chat, Team, Attività, Calendario, Chiamate, File, and three dots for more options. Below the navigation bar, there's a message from Marco Marcobelo Balletti: "bello cliente confronti di prestazioni tra cose e shahhaha". A message from Valeria Cardellini says: "La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione." On the right side, there's a "Chat della riunione" window titled "Riunione" with the message "Registrazione iniziata".

Queries for the team

- 1 student in the team: queries 1 and 2; data ingestion is optional
- 2 students in the team: all the three queries plus data ingestion
- 3 students in the team: all the three queries plus data ingestion and optional part

A presentation slide titled "V. Cardellini, F. Rossi - SABD 2019/2020". The slide has a dark background with a grid of icons. At the top left, it says "Diapositiva 10 di 11". At the top right, there are buttons for "Assumi il controllo" and "Torna al relatore". The slide features a large circular icon with initials "VC" and several smaller circular icons with user profiles. At the bottom, there are buttons for "App" and "Guarda".

Data ingestion

- Which framework to ingest data into HDFS?
 - Flume, Kafka, NIFI, ...
 - Which format to store data?
 - csv, columnar format (Parquet), row format (Avro), ...
 - Where to export your results?
 - HBase, Redis, Kafka, ...

