

The screenshot shows a Microsoft Teams meeting window. At the top, there's a dark header with the Microsoft Teams logo, file navigation (Modifica, Visualizza, Finestra), a search bar ("Esegui una ricerca o digita un comando"), and system status (Lun 4 mag 12:29:22 Francesco Marino). On the far right, there's a user profile icon. The main content area has a yellow border and displays a presentation slide. The slide features the logo of "TOR VERGATA UNIVERSITÀ DEGLI STUDI DI ROMA" (a green stylized 'U' icon) and text: "Macroarea di Ingegneria Dipartimento di Ingegneria Civile e Ingegneria Informatica". Below this, the title "Project 1" is displayed in large blue letters. The main text on the slide reads: "Corso di Sistemi e Architetture per Big Data" in bold black font, followed by "A.A. 2019/2020" and the names "Valeria Cardellini, Fabiana Rossi". At the bottom of the slide, there are navigation controls: "Diapositiva 1 di 11", "Assumi il controllo" (with a red '6' badge), "Torna al relatore", and a message "Il microfono è disattivato.". The bottom of the screen shows the Teams interface: a participant list with icons for Valeria Cardellini, luigi corsi, ilenia rocca, marco ballotti, and Valeria Cardellini; a toolbar with icons for +6, DN, AC, and other participants; and a bottom dock with various application icons (Safari, Mail, Chrome, etc.). On the right side, there's a "Chat della riunione" panel showing a message about a recording starting, and a "Rispondi" button with a reply icon.

A screenshot of a Microsoft Teams meeting interface. The main content is a slide titled "Project delivery" with a list of requirements. The requirements include: Submission deadline (May 29, 2020), Your presentation (June 4, 2020), What to deliver (Link to cloud storage, Project report in ACM or IEEE format, Presentation slides max. 15 minutes), and Team (2 students per team, also possible 1 student). The slide footer shows "V. Cardellini, F. Rossi - SABD 2019/2020". The Teams navigation bar on the left includes icons for Azioni, Chat, Team, Attività, Calendario, Chiamate, File, and more. A message at the top says "La registrazione è stata avviata." (Recording has started). The bottom of the screen shows the Mac OS Dock with various application icons.

A screenshot of a Microsoft Teams meeting interface. The main content area displays a slide with the title "Dataset" and a bulleted list: "You will use two real datasets on Covid-19", "Both datasets are available in CSV format", and "Datasets are available at" followed by two GitHub links. The bottom of the slide shows navigation controls: "Diapositiva 3 di 11", "Assumi il controllo", and "Torna al relatore". The footer of the slide includes the names "Valeria Cardellini" and "V. Cardellini, F. Rossi - SABD 2019/2020". A toolbar below the slide contains icons for video, audio, and other controls. The Teams sidebar on the left shows various team channels like "Azioni", "Chat", "Team", "Attività", "Calendario", "Chiamate", and "File". A notification bar at the top indicates that recording has started. The bottom dock features a row of icons for various Mac OS applications.

The screenshot shows a Microsoft Teams meeting interface. The main content is a slide titled "Dataset 1: schema" with a bulleted list of requirements for the dataset. A callout box highlights that "Most data are cumulative". The bottom of the slide shows the presentation navigation bar with controls for back, forward, and search, along with the presenters' names: Valeria Cardellini, V. Cardellini, F. Rossi - SABD 2019/2020, and the slide number 3. The Teams ribbon on the left includes icons for Azioni, Chat, Team, Attività, Calendario, Chiamate, File, and more. A message at the top indicates that recording has started. On the right, there's a "Chat della riunione" pane showing a message about the recording starting, and a "Rispondi" pane showing a reply from another participant.

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

Most data are **cumulative**

Diapositiva 4 di 11 | Assumi il controllo | Torna al relatore | 6 | 3 | Rispondi

Valeria Cardellini | V. Cardellini, F. Rossi - SABD 2019/2020 | 3 | Rispondi

Azioni | Chat | Team | Attività | Calendario | Chiamate | File | ... | Informativa sulla privacy | Ignora | Chat della riunione | Riunione | Registrazione iniziata | Rispondi | Rispondi

A screenshot of a Microsoft Teams meeting interface. The top navigation bar shows 'Microsoft Teams' and the date 'Lun 4 mag 12:29:41'. A search bar says 'Esegui una ricerca o digita un comando'. On the left, a sidebar has icons for Azioni, Chat, Team, Attività, Calendario, Chiamate, File, and three dots. A message at the top says 'La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.' To the right, there's an 'Informativa sulla privacy' button and an 'Ignora' button. The main content area has a dark background with a yellow border. It features a title 'Dataset 2: schema' and a bulleted list: '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'. Below the list are controls for navigating slides ('Diapositiva 5 di 11'), taking control ('Assumi il controllo'), and returning to the speaker ('Torna al relatore'). There are also video controls (camera, microphone, up arrow, three dots, document, file, red phone), a slide number '4', and participant names: Valeria Cardellini, V. Cardellini, F. Rossi - SABD 2019/2020, luigi corsi, ilenia rocca, marco balletti, and Valeria Cardellini. At the bottom, there's a taskbar with various icons like smiley face, rocket, calendar, gear, etc., and a floating window on the right titled 'Rispondi' showing the presentation slide.

Microsoft Teams Modifica Visualizza Finestra ?

Esegui una ricerca o digita un comando

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

Informativa sulla privacy

Federica Sticker

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

Query 1

Using dataset 1, for each week determine the average number of cured people and the average number of swabs

< Diapositiva 6 di 11 > Assumi il controllo Torna al relatore 6

26:22

Valeria Cardellini V. Cardellini, F. Rossi - SABD 2019/2020

5

+6 DN AC ilia corsi ilenia rocca marco balletti Valeria Cardellini

Rispondi

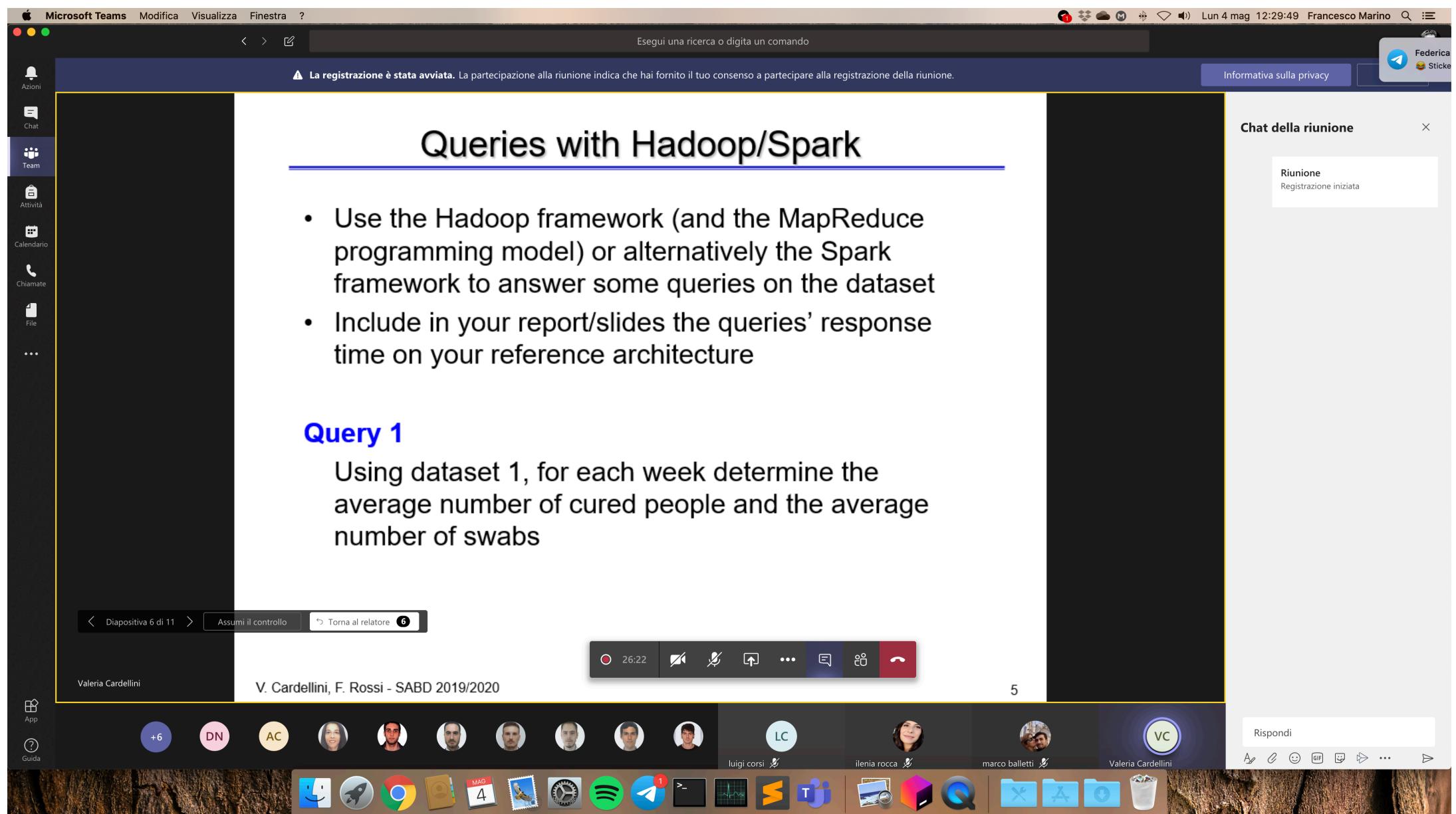
App Guida

Microsoft Teams

Lun 4 mag 12:29:49 Francesco Marino

Chat della riunione

Riunione Registrazione iniziata



Microsoft Teams Modifica Visualizza Finestra ?

Esegui una ricerca o digita un comando

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

Informativa sulla privacy Ignora

Queries with Hadoop/Spark

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

< Diapositiva 7 di 11 > Assumi il controllo Torna al relatore 7

31:56

Valeria Cardellini V. Cardellini, F. Rossi - SABD 2019/2020

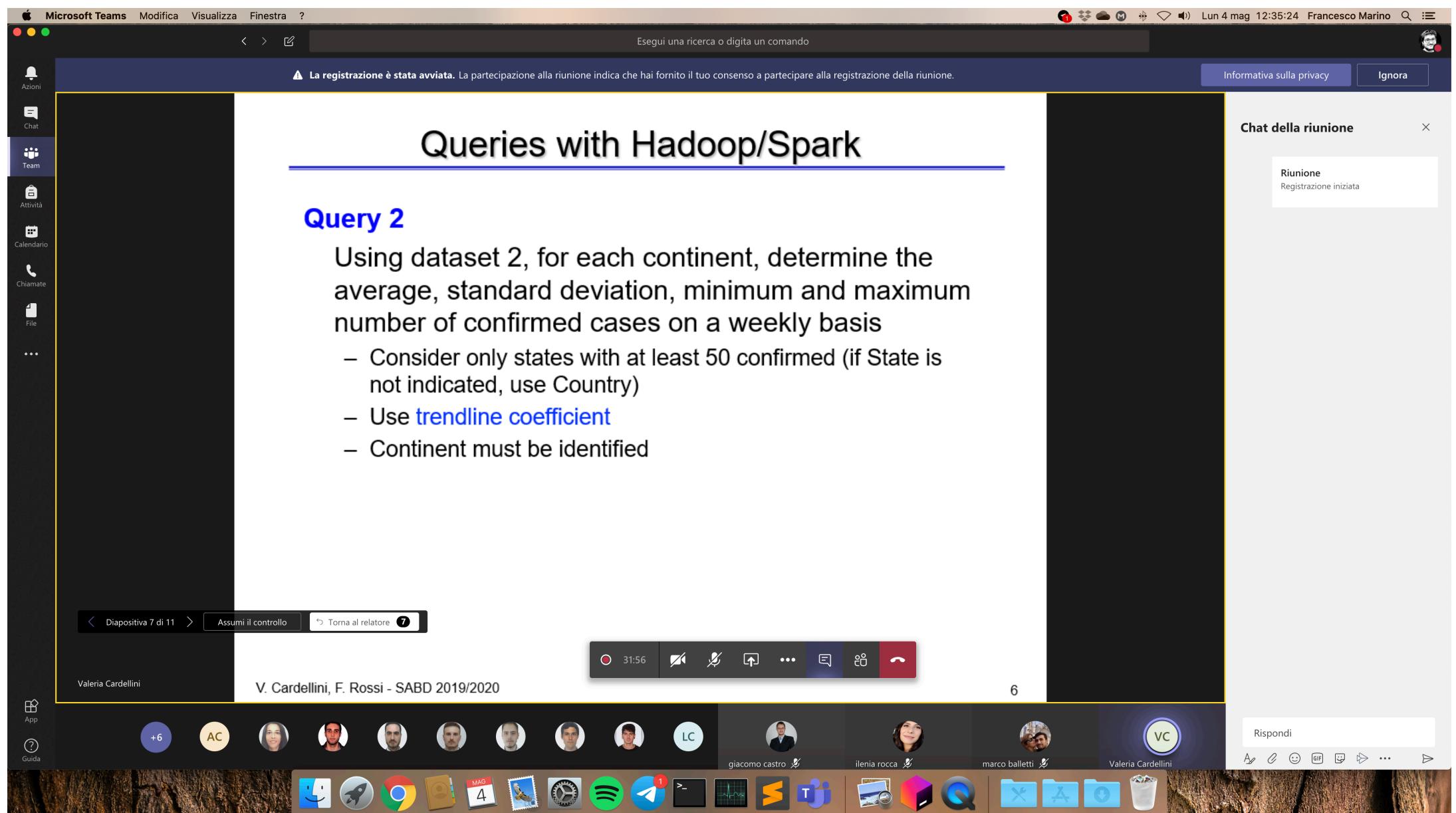
6

Rispondi

App Guida

+6 AC giacomo castro ilenia rocca marco balletti Valeria Cardellini

VC



Microsoft Teams Modifica Visualizza Finestra ?

Esegui una ricerca o digita un comando

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

Informativa sulla privacy Ignora

Queries with Hadoop/Spark

Query 3

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

- Compare the performance of a naïve implementation of K-means with that provided by Spark MLlib or Apache Mahout

< Diapositiva 8 di 11 > Assumi il controllo Torna al relatore 9

41:58

Valeria Cardellini V. Cardellini, F. Rossi - SABD 2019/2020

7

Rispondi

A C Gif ...

App Guida

+6 AC giacomo castro ilenia rocca marco balletti Valeria Cardellini

7

The screenshot captures a Microsoft Teams meeting in progress. The main content area displays a presentation slide titled "Queries with Hadoop/Spark" and a sub-section "Query 3". The slide text describes using dataset 2 and top-50 states to perform K-means clustering with K=5. Below the slide, a list item suggests comparing a naive implementation with Spark MLlib or Apache Mahout. At the bottom of the slide, there are navigation controls for the presentation (back, forward, search), a timer (41:58), and a control bar for the meeting (Assumi il controllo, Torna al relatore). The Teams interface also shows participant names (Valeria Cardellini, V. Cardellini, F. Rossi - SABD 2019/2020), a list of participants (giacomo castro, ilenia rocca, marco balletti, Valeria Cardellini), and a toolbar with various application icons (e.g., Mail, Calendar, File, etc.). A status bar at the bottom shows the date (Lun 4 mag 12:45:25) and the user's name (Francesco Marino).

Microsoft Teams Modifica Visualizza Finestra ?

Esegui una ricerca o digita un comando

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

Informativa sulla privacy Ignora

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

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

Valeria Cardellini V. Cardellini, F. Rossi - SABD 2019/2020

43:08

giacomo castro ilenia rocca marco ballesti Valeria Cardellini

Rispondi

8

App Guida

The screenshot captures a Microsoft Teams meeting in progress. The main content area displays a presentation slide with the title "Optional part" and a bulleted list of instructions for a team of three students. The list involves using Hive/Pig or Spark SQL to address queries and comparing performance with a pure Hadoop/Spark-based solution. Navigation controls at the bottom left allow switching between slides and taking over the presentation. The bottom of the screen features a dock with icons for various Mac OS applications like Finder, Mail, and Safari. A sidebar on the right shows a "Chat della riunione" (Meeting Chat) with a message about recording starting. The Teams interface includes standard controls for audio, video, and sharing at the bottom center.

The screenshot shows a Microsoft Teams meeting interface. The main content is a slide titled "Queries for the team" with the following bullet points:

- 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

At the bottom of the slide, there are navigation controls: "Diapositiva 10 di 11", "Assumi il controllo", and "Torna al relatore". A timer shows 46:00. Below the slide, the footer displays the names "Valeria Cardellini" and "V. Cardellini, F. Rossi - SABD 2019/2020". The bottom bar shows participant icons for "giacomo castro", "ilenia rocca", "marco balletti", and "Valeria Cardellini". The Teams ribbon on the left includes icons for Azioni, Chat, Team, Attività, Calendario, Chiamate, File, and others.

Microsoft Teams Modifica Visualizza Finestra ?

Esegui una ricerca o digita un comando

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

Informativa sulla privacy Ignora

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, ...

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

46:06

Valeria Cardellini V. Cardellini, F. Rossi - SABD 2019/2020

10

Rispondi

A C Gif ...

App Guida

+6 AC giacomo castro ilenia rocca marco ballesti Valeria Cardellini

12 4 MAO 3 S T Microsoft Teams

