CASE-STUDY (NUMBER - 6) – SOLUTION SUBMISSION

ON

AZURE ANALYTICS

BY

NAME –SUPRIYA BHARATHA ROLL NUMBER DXCAB1211

BATCH:DXC-262-ANALYTICS-B12-AZURE COMPANY – DXC TECHNOLOGY

TRAINING UNDER – MANIPAL PRO LEARN TRAINER NAME – MR. AJAY KUMAR DATE OF SUBMISSION – 6th JUNE 2022 TOTAL NUMBER OF QUESTIONS: 09 EMPLOYEE DOMAIN - AZURE ANALYTICS

* THE QUESTIONS:

1. Explain what is in-Memory computation in details?

2. Explain advantages of Spark framework ?

3. Explain components of Spark with block diagram ?

4. Explain benifits of in-Memory computation ?

5. Explain major difference between Hadoop & Spark ?

6. Explain features of Spark?

7. Write a Py-Spark program to create Dataframe from RDD & explain with screenshots

& steps ?

8. Explain what is RDD & why it is needed ?

9. Write a Py-Spark program to make the column in Upper case & explain with screenshots & steps ?

1. Explain what is in-Memory computation in details?

SOLUTION:

* In-memory Computing Processing in memory is one approach to overcoming the von Neumann bottleneck, which is a limitation on throughput caused by the latency inherent in the standard computer architecture.
* In-memory computing primarily relies on keeping data in a server's RAM as a means of processing at faster speeds.
* In-memory computing especially applies to processing problems that require extensive access to data-analytics, reporting or data warehousing, and big data applications.

1. Explain advantages of Spark framework ?

SOLUTION:

* Speed.
* Ease of Use.
* Advanced Analytics.
* Dynamic in Nature.
* Apache Spark is powerful.
* Increased access to Big data.

1. Explain components of Spark with block diagram ?

SOLUTION:

1. Spark SQL Structured data : Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.
2. Spark Streaming Real-time : Spark Streaming is an extension of the core Spark API that allows data engineers and data scientists to process real-time data from various sources
3. Milb Machine Learning : MLlib is a low-level machine learning library. It can be called from Java, Scala and Python programming languages. It is simple to use, scalable and can be easily integrated with other tools and frameworks.
4. GraphX Graph Processing : GraphX is useful in giving overall information about the graph network like it can tell how many triangles appear in the graph and apply the PageRank algorithm to it.

**Spark Streaming Real-time**

**Milb Machine Learning**

**GraphX Graph Processing**

**Spark SQL Structured data**

**Spark Core**

1. Explain benifits of in-Memory computation ?

SOLUTION:

* Benefits of In-memory computing
* Better, faster, decision making.
* Ability to reduce cost.
* Identify competitive opportunities.
* Grow revenue.
* More efficient application.
* Reduce risk.
* It's best suited for performing real-time analytics, and developing and deploying real-time applications.
* In-memory Computing Imperative:
* Avoid movement of detailed data.
* Calculate first, then move the results.

1. Explain major difference between Hadoop & Spark ?

SOLUTION:

* Spark was built on the top of the Hadoop MapReduce, Spark process the data much quicker than Hadoop framework.
* Hadoop 1x times faster - Spark will be 100x times faster.
* Apache Hadoop provides batch processing, Apache Spark provides both batch processing and stream processing.
* Hadoop is difficult to use, Spark is easier to use.

1. Explain features of Spark?

SOLUTION:

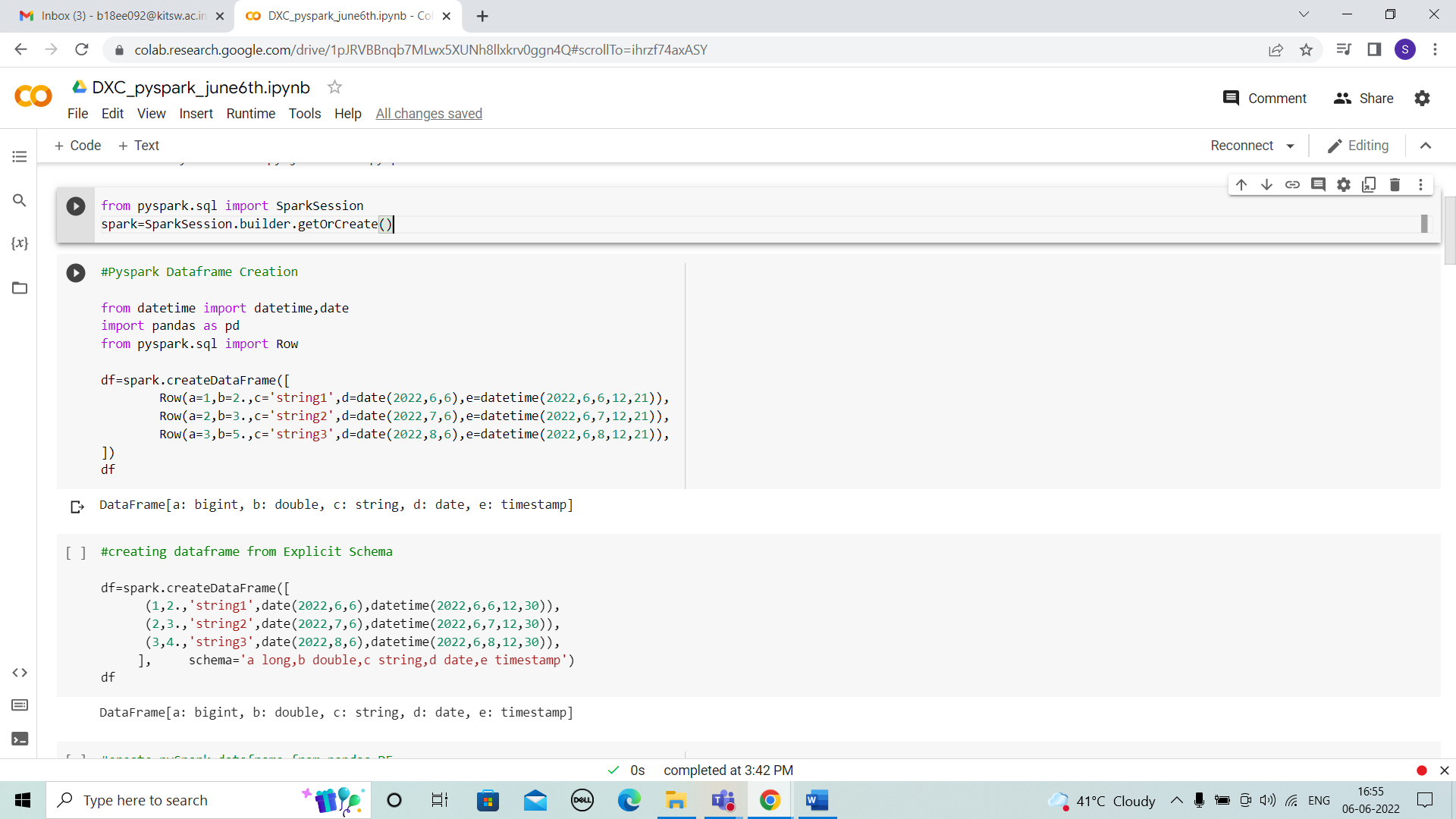
**Features of Apache Spark**

* Fast :It provides high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.
* Easy to Use: It supports various languages like Java, Python, Scala, Sql, R, It facilitates to write the application in Java, Scala, Python, R, and SQL. It also provides more than 80 high-level operators.
* Supports Various Libraries: It provides a collection of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming.
* Supports Realtime Streaming:
* Lightweight : It is a light unified analytics engine which is used for large scale data processing. Runs Everywhere - It can easily run on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud.

1. Write a Py-Spark program to create Dataframe from RDD & explain with screenshots & steps ?

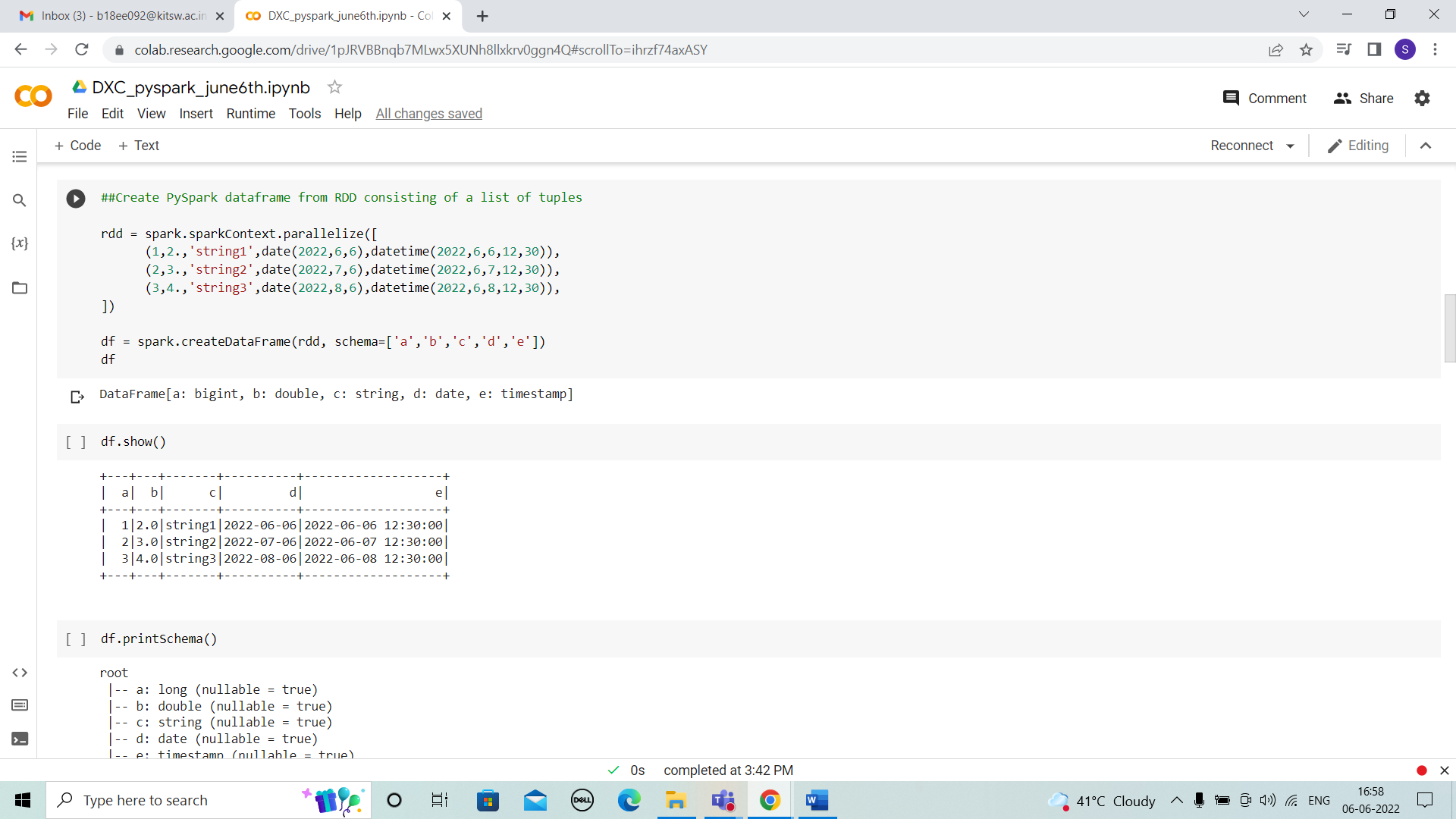
SOLUTION:

STEP 1: Create PySpark RDD





STEP 2: In PySpark, when you have data in a list meaning you have a collection of data in a PySpark driver memory when you create an RDD, this collection is going to be parallelized.





1. Explain what is RDD & why it is needed ?

SOLUTION: RDD - Resilient Distributed Dataset:

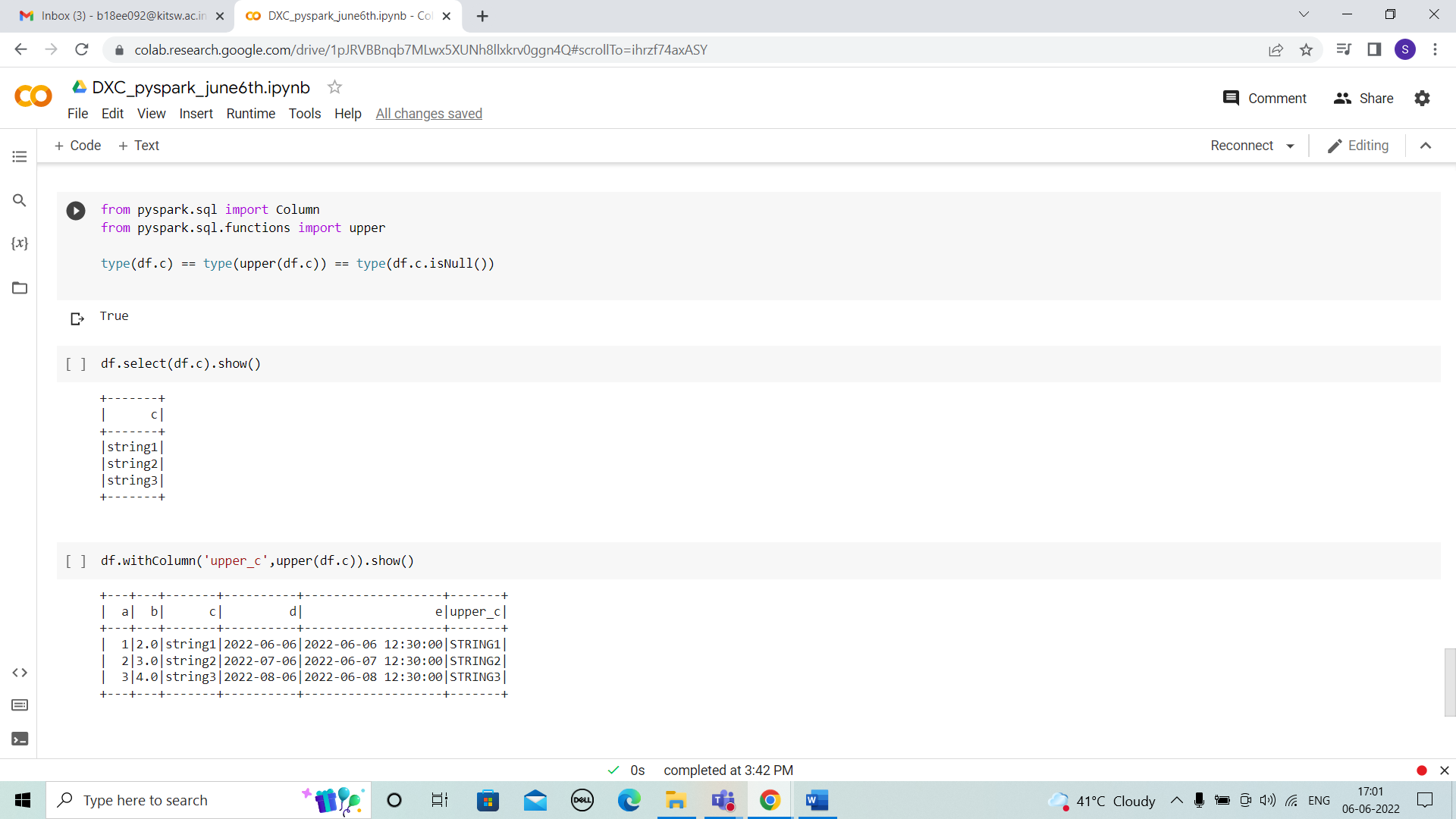
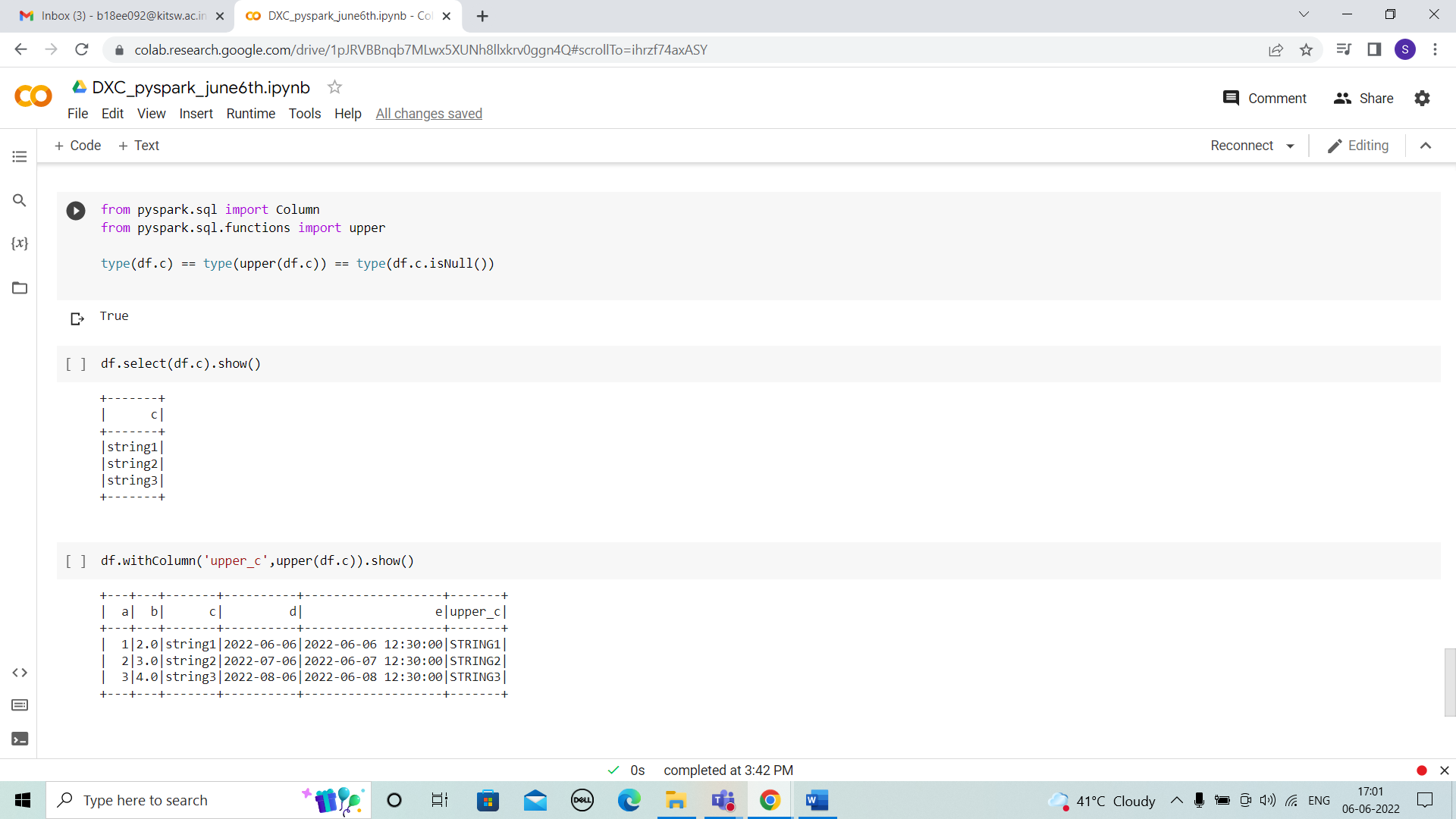
* It is basis building block of Spark,
* The RDD (Resilient Distributed Dataset) is the Spark's core abstraction.
* It is a collection of elements, partitioned across the nodes of the cluster so that we can execute various parallel operations on it.
* There are two ways to create RDDs:
* Parallelizing an existing data in the driver program
* Referencing a dataset in an external storage system, such as a shared filesystem, HDFS, HBase, or any data source offering a Hadoop InputFormat.
* it supports in-memory processing computation. This means, it stores the state of memory as an object across the jobs and the object is sharable between those jobs.

1. Write a Py-Spark program to make the column in Upper case & explain with screenshots & steps ?



SOLUTION:

STEP 1: In order to convert a column to Upper case in pyspark we will be using upper() function.





RESULTS:

I HAVE SUCCESSFULLY ANSWERED ALL THE QUESTIONS AS PER ASSIGNMENT REQUIREMENT.

CONCLUSIONS:

All the questions have been solved successfully with all the concepts that have been covered in the training session. It’s really a great experience of learning while solving the cases. This assignment gave me immense confidence regarding my ability to upskill in new technologies.