## Week 2

1. Imagine you are given access to a data table that contains millions of rows. Hundreds of thousands of records are added every hour. Which characteristics of big data does this scenario relate to?
   1. Velocity and volume.
2. What are examples of unstructured data?
   1. Social media posts.
   2. Videos.
3. What are three of the high-level big data needs identified for organisations?
   1. Collecting and storing data, and generating data insights.
4. Whilst spreadsheets are useful for working with data in general, they aren’t robust enough to work with big data.
   1. True.
5. What’s true about Apache Spark?
   1. Apache Spark can connect to many different data sources and work with a wide variety of data types.
   2. Apache Spark processes a single set of data to be processed on multiple machines.
6. What do we call the group of machines that Spark uses to power queries?
   1. Clusters.

## Week 3

1. What is a Databricks table?
   1. A collection of structured data.
2. You can show a sampling of a table by running a SELECT \* FROM table\_name query. What’s one other way you can preview the data in the table?
   1. Select it from the Data tab in your Databricks workspace.
   2. Run the command SLECT \* FROM table\_name LIMIT 5.
3. Which statement about a table’s schema is true?
   1. It describes the structure of your table.
4. What’s wrong with the following query?

SELECT birthyear,

ROUND(AVG(salary), 2) AS avgSalary

FROM

peopleSavings

* 1. It’s missing a GROUP BY clause.

1. How is a VIEW different from a TABLE?
   1. Views won’t appear for preview from the Data tab.
   2. Tables are accessible in your workspace until you decide to drop them. Views are automatically dropped.

### Week 4

1. What is the first action Spark’s SQL optimiser performs before processing your query?
   1. It checks the query against the metadata catalogue to verify that the table and columns you’re referencing actually exist.
2. Spark SQL can auto-optimise your query by reordering the sequence of commands so that it can process the data in the most efficient way. What is that stage in the optimisation process called?
   1. Logical optimisation.
3. A cluster is a set of computation resources that power your query. Within a cluster, a driver breaks up and distributes work to the workers. What happens if one of the workers shuts down?
   1. Nothing really; the work will just be redistributed among the workers that are continuing to run.
4. The Spark UI contains a lot of visualisations about how your data is being processed and exactly what’s going on within your cluster. In which area of the UI would you find a list of all the SQL queries that have been performed since your cluster started running?
   1. The SQL tab.
5. If you see that a column in the table you’re working with beings with a “p\_”, what does it likely indicate about the table?
   1. It is partitioned by that column.
6. What is Spark caching?
   1. It stores data in temporary storage across machines in your cluster.
7. What does it mean to partition a table?
   1. It allows you to write sub-directories for the table in your file system.

## Week 6

1. What statement best describes a higher-order function?
   1. Higher-order functions allow you to define functions that manipulate arrays in SQL.
2. Which statement best describes the purpose of the iterator in a higher-order function?
   1. The iterator is a place-holder variable that you use within the function definition. When you write the function, you should think of the iterator as a single element of the array you’re working with.
3. Which statement best describes a sub-query?
   1. A sub-query is a query within a query.
4. Imaging you’re applying the reduce function to an ARRAY type column and you’re trying to find the mean of each array. This is how you’ve written the REDUCE function:  
   REDUCE(demo\_array, 0, (element, acc) -> element + acc, acc -> (acc div size(demo\_array)))

Which statement best describes the variable acc?

* 1. It is the accumulator variable; it holds a running total of the elements as they are summed in the first function of this command.

1. What statement best describes the difference between Rollup and Cube operators.
   1. Rollup generates subtotals and grand totals for a set of columns. Cube generates subtotals and grand totals for all possible combinations of the grouping columns specified.
2. Identify the error in the query below:  
   SELECT  
    device\_type,  
    ROUND(AVG(avg\_daily\_temp\_c), 4) AS avgTemp,  
    ROUND(STD(avg\_daily\_temp\_c), 4) AS stdTemp,  
   FROM AvgTemps  
   WHERE device\_type = getArgument()  
   GROUP BY device\_type
   1. The get\_argument function needs to include the name of your widget as a parameter.
3. What defines the “window” in a window function?
   1. The OVER clause.

## Week 5

1. What is a reason for including the DROP TABLE command **before** the table creation statement as shown below?

DROP TABLE IF EXISTS DCDataRaw;  
CREATE TABLE DCDataRaw  
USING parquet  
OPTIONS (  
 PATH “/mnt/training/iot-devices/data-centres/2019-q2-q3”  
 )

* 1. It ensures that we can run the notebook more than once without encountering an error due to a naming conflict.

1. What additional information is provided when you use a DESCRIBE EXTENDED command?
   1. Detailed table information, like which database holds the data and the table name.
2. What are 3 complex data types you can work with using Spark SQL?
   1. ARRAY, MAP and STRUCT.
3. What keyword always starts a Common Table Expression (CTE)?
   1. WITH.
4. How is data stored in a MAP type column?
   1. Key-value pairs.
5. Identify the error:

SELECT key, dc\_id, date, value.description, value.ip, value.temps, value.co2\_level

1. FROM  
    ExplodeSource  
   WITH ExplodeSource  
   AS  
   (  
    SELECT dc\_id, to\_date(date) AS date, EXPLODE(source)  
    FROM  
    DCDataRaw  
   );
   1. The WITH clause must go at the start of a CTE.
2. If you want to create a new table based on an existing table, which design pattern could you use?
   1. Common Table AS Select (CTAS).
3. When working with arrays, we can access each element by its position on the array. We call that position the element’s
   1. Index.
4. Imagine that you used a CTE to access some nested data. Later on in the notebook, you want to reference that table or view that was created by the CTE, but you get an error message when you write the query. What is the most likely cause of the error?
   1. CTEs do not create tables or views. Queries must be performed directly after definition.