# AC51011: Big Data Analysis
# AC41011, AC51047 : Advanced Big Data Analysis
# Big Data Storage and Data Warehousing Assignment

**Assignment Description:** Given are three problems related to the use of NoSQL databases and data warehousing in storing big data sets, as well as performing some basic analysis on these sets. Write up a report that will contain solutions to all three problems. For each problem, write **how** you have solved it (answering also all the questions that can be found in the description of the problem) and what was the reasoning that you have employed in solving the problem, attaching also screenshots of your solution where relevant (i.e. where you used graphical user interface to solve parts of the problem). You will also need to attach the actual code that you have used (in particular, if you have used command line/terminal interface, and where you need to write SQL queries). Note that if you **only** write the code and attach the screenshots of the solutions for the problems, without any explanation, you will be severely penalised in terms of the actual grade that you will obtain.

**How Will Your Submission Be Graded:** Grades will be determined by evaluating the overall quality of the solution, together with the quality of the report. The grade will be between **AB** and **A1**, and it will comprise 20% of your final grade for the module. Note that solving all the problems and answering all the questions specified, together with a good report, will give you the grade A5. To get a better grade, you will need to present an especially elegant/good solution, or you will have to do extra work (of your own choice) in addition to the specification.

**What We are Looking for in the Report:** Good explanation of the reasoning that led you to the solution, as well as of any design decisions that you have made. Good explanation of queries that you wrote, good reasoning about the tables/cubes that you have created and choice of the table/cube features (primary keys of tables, dimensions and measures in the cube etc.).

**Dataset Description:** The dataset for this assignment is stored in the store.csv file that can be found on MyDundee page of the assignment. It is a dataset of transactions from a store that sells different kind of stationery and hardware. The dataset contains a variety of information about each transaction, including information about the product (and its category and subcategory), customer (including customer name and the customer segment), shipping mode, store where the product was sold (including region, state and city where the store is located), the amount of money in sales and so on.

## Problems

1. **S3 Storage Service** (**10% of the grade**). Create a new S3 bucket and put in the store.csv file into that bucket.

2. **Key-Value Stores** (**40% of the grade**). Using Amazon DynamoDB or Cassandra, create a new table that will model the data from the input dataset and load the data from the S3 bucket that you created in problem 1 into the newly created table. Attach screenshots if you have done this using the DynamoDB Dashboard interface, or the actual command(s) for the creation of the table if you have used the AWS CLI interface form the AWS terminal.

    a. (**10%**) What did you put as a partition key in this table and why? What (if anything) did you put as a sort key and why? What other attributes could have you put for partition and sort key, compared to your solution?

    b. (**10%**) Give at least two examples of queries that could be efficiently executed on your table. Describe the queries both using English language and the PartiQL language. Also describe how you would perform these queries on the table that you have created (attach screenshots if you have used the DynamoDB Dashboard interface, or attach the commands if you have used the AWS CLI interface from the AWS terminal)

    **Hint:** To know whether a query can be 'efficiently executed' or not, you can try running your query in PartiQL editor on AWS DynamoDB service, and observe whether you get all the results or just first 200 results, and in what time.

    c. (**10%**) Give at least two examples of queries that could **not** be efficiently executed on your table. Describe the queries both using the English language and the PartiQL language. Explain also **why** the queries in this part could not be efficiently executed.

    **Hint:** Again, to judge whether a query can be 'efficiently executed' or not, try running a query and look at the results (and the information about the result) in the PartiQL editor.

    d. (**10%**) What would you do to make the execution of the queries for the part c more efficient? How would this make the execution of the queries faster? Attach screenshots if you have used the DynamoDB Dashboard, or the actual commands performed if you have used the AWS CLI interface from the AWS terminal.

3. **Data Warehousing** (**50% of the grade**). Using Amazon Redshift or Apache Hive, create a new external table that will serve as a wrapper for the DynamoDB/Cassandra table in the

previous exercise, link the table to the DynamoDB/Cassandra table.

a. (**20%**) Write SQL queries on the table to answer the following questions:

   i. How many different product categories do there exist in the dataset?
   ii. How many orders were there in each of the regions?
   iii. Which product category has the highest total sales amount?
   iv. Which states have recorded the biggest and the smallest profit?
   v. Which city in the most profitable state has the least profit?
   vi. Which product subcategory had the most orders that were of the critical priority?
   vii. What is the total amount of sales for 'Small Business' customer segment?
   viii. What product sub-category was the most often purchased in the 'Corporate' customer segment, and what is the total amount of sales in all of these transactions?
   ix. What is the total profit for all of the products of the 'Office Machines' subcategory in the 'California' state?

b. (**10%**) Think of a data cube that would help you answer the questions iv), vii), viii) and ix). Write an SQL query that would create such a cube, and also queries on it to answer these three questions, as well as the following two:

   i. What is the total profit achieved over all the transactions?
   ii. How many transactions were there in the Minnesota state?

c. (**10%**) Create another cube for *roll up* and *drill down* operations that would help you answer the questions ii), iv) and v). Write an SQL query that would create this cube.

d. (**10%**) Explain what the benefit is of creating data cubes in parts b) and c) and why are the queries on these data cubes more efficient than queries on the top-level table, as in the part a). What are the drawbacks of creating these cubes?