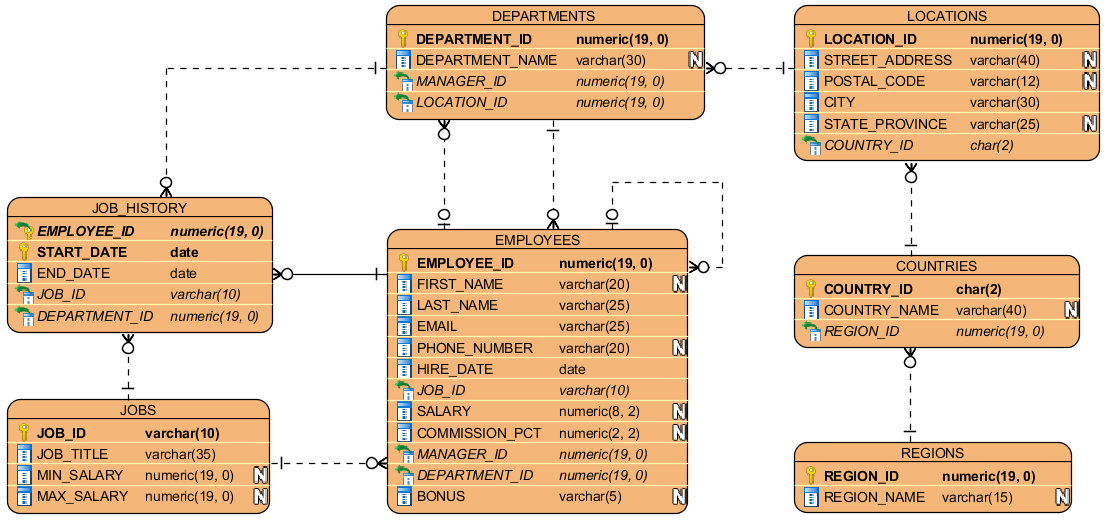
# Topic 4 – SQL Joins

## Familiarise yourself with the Human Resource (HR) schema from Oracle

Whereas in the lectures we use a small database consisting of only three table, for the labs we will be using the demo tables (schema) supplied by Oracle. The schema describes a human resources database used by a large multi-national company.

Use the Entity Relationship Diagram (ERD) below to familiarise yourself with the tables, relationships and attributes.



# Worked Examples

### Q1. List the names of the employees and the names of the departments they work in.

If we examine the schema above we see that the employee names are in one table (employees) but the department names are in a different table (departments), we are therefore being asked to get information from two different tables and present them as if they were in one table. This is what joins are for.

We start by identifying the columns we need and which tables they are in – first\_name, last\_name in the employees table and department\_name from the departments table.

Next, we have to identify how to join the tables. In this case, we know we need an inner join because for every row in the employees table we want to find a single row in the departments table that matches it. We need to provide the condition for the match – i.e. how will the database know when a row from the departments table matches the row from the employees table. Looking at the schema again we see that both have a column called department\_id. The join condition is that the row in the departments table has the same value for department\_id as the row in the employees table.

We can therefore write the required SQL query:

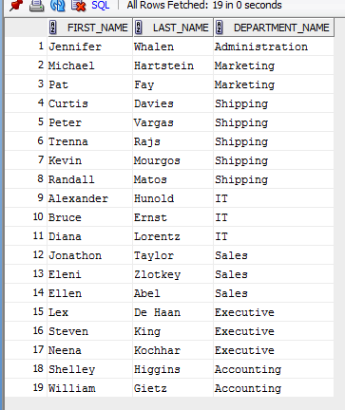
SELECT first\_name, last\_name, department\_name

FROM employees

JOIN departments

ON employees.department\_id = departments.department\_id;

This gives the following result:



Notice that we only have 19 records here whereas “SELECT \* FROM employees;” shows 20 records. This is because the record for Kimberley Grant has NULL for the department\_id which cannot match anything in the departments table and so no record is returned for her.

**NEW MATERIAL**

We say that we are “joining on” the department\_id column because that is the column used to match rows. It might be useful, therefore, to include that in the output. Try the following SQL:

SELECT first\_name, last\_name, department\_name, department\_id

FROM employees

JOIN departments

ON employees.department\_id = departments.department\_id;

This throws an error: “column ambiguously defined”. The error means that one of the columns can have more than one meaning – i.e. the database cannot determine which column you are actually referring to. This is because there are two columns called department\_id involved in the query and we have not specified which one we mean.

The solution is to specify not only the column name but also the table name:

SELECT first\_name, last\_name, department\_name,

**employees.department\_id**

FROM employees

JOIN departments

ON employees.department\_id = departments.department\_id;

Notice that the error will be thrown if the table name is missed in the ON clause as well.

Try the following SQL:

SELECT first\_name, last\_name, department\_name,

employees.department\_id

FROM employees

JOIN departments

ON employees.department\_id = **department\_id;**

You will again get the same error.

Rather unfortunately, the error message does not say which column is ambiguously defined. (Compare to the error message you get when trying something like SELECT department\_nme FROM departments; where the error tells you which column name is invalid). However, the error message does report something like “Error at Line: 1 Column 33”. This is telling you where in your SQL the error is. Do not confuse this column with a column of a table. This is referring to the worksheet you have open and you are typing in. Column 33 means that the error starts at the 33rd character on that line.

The editor window you have open has line numbers on the left hand side so finding the line in your worksheet where the error occurred is straightforward. Finding the column (meaning the number of characters along the line) is less easy. Counting can be difficult because you end up dealing with reasonably large numbers.

Look at the bottom right of the SQL developer window – you should see a light blue band with white writing on. The left-most part of that gives the “Line” and “Column” and now move the cursor around inside the worksheet (use the arrow keys) and notice how the Line and Column change. So to find the column where the error occurred you can start moving along the line keeping an eye on the Column counter in the bottom right of the screen.



This way we can use the error message to identify which column in the SELECT clause is ambiguous. The column will be the one that starts at the Line and Column reported in the error message.

**IMPORT NOTE**

We mentioned in the lecture about the dangers of the Natural Join. This question is a good example. We might be tempted to use the Natural Join here.

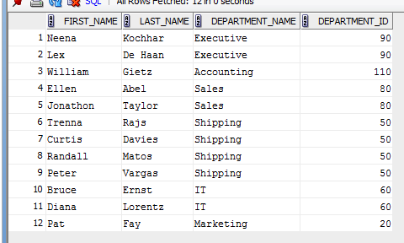
Try the following SQL:

SELECT first\_name, last\_name, department\_name, department\_id

FROM employees

NATURAL JOIN departments;

The results should be:



This is different to the result we got when using the more fully-specified INNER JOIN … ON method. The reason is that the employees and departments tables don’t just share a column called department\_id. They also both have a column called manager\_id. Therefore, the Natural Join will join on both of these columns. The equivalent SQL for the Natural Join of these two tables is:

SELECT first\_name, last\_name, department\_name,

employees.department\_id

FROM employees

JOIN departments

ON employees.department\_id = departments.department\_id

**AND employees.manager\_id = departments.manager\_id;**

Notice the additional join condition which is what causes the results to be different. This mistake can happen if you use a natural join but is impossible using the INNER JOIN … ON method. Even if you were very careful when using the natural join, it is easy to imagine a situation in which the departments table did not originally have a manager\_id in it and so the two queries were identical. Then, at some later date, the manager\_id column was added to the departments table and now, without changing the SQL of the query at all, the results change and we are returning the wrong information.

This is why you should be very careful with and avoid using the natural join.

Another, widely used, alternative is to write the query as:

SELECT first\_name, last\_name, department\_name,

employees.department\_id

FROM employees, departments

WHERE employees.department\_id = departments.department\_id;

This is the older, traditional style required before the SQL standard added the INNER JOIN … ON keywords. Conceptually it is performing a cross join (which returns every possible combination of rows) and then filters the result using the WHERE clause to only return the combinations that match. Most modern database systems will simply translate this into an inner join and so the two syntaxes are effectively equivalent.

The choice of which to use depends largely on the number of tables involved in the join because the ultimate consideration is readability. When there are only a small number (e.g. fewer than five) it is readable to use the INNER JOIN … ON style and that style makes it explicit that we’re doing an inner join and what the condition is. The traditional style relies on people spotting the comma.

On the other hand, when there are a lot of tables involved in the join all of the INNER JOIN … ON clauses can become difficult to spot and identify. In this case a comma separated list of table names may be clearer with a collection of WHERE clauses underneath. Compare, for example:

SELECT col1, col2, col3, col4, col5, col6, col7, col8

FROM tbl1,

tbl2,

tbl3,

tbl4,

tbl5,

tbl6,

tbl7

WHERE tbl1.col2 = tbl2.col2

AND tbl2.col3 = tbl3.col3

AND tbl3.col4 = tbl4.col4

AND tbl4.col5 = tbl5.col5

AND tbl5.col6 = tbl5.col6

AND tbl6.col7 = tbl7.col7

AND tbl7.col8 = ‘something’;

With

SELECT col1, col2, col3, col4, col5, col6, col7, col8

FROM tbl1

INNER JOIN tbl2 ON tbl1.col2 = tbl2.col2

INNER JOIN tbl3 ON tbl2.col3 = tbl3.col3

INNER JOIN tbl3 ON tbl2.col3 = tbl3.col3

INNER JOIN tbl4 ON tbl3.col4 = tbl4.col4

INNER JOIN tbl5 ON tbl4.col5 = tbl5.col5

INNER JOIN tbl6 ON tbl5.col6 = tbl5.col6

INNER JOIN tbl7 ON tbl6.col7 = tbl7.col7

WHERE tbl7.col8 = ‘something’;

Whichever style you choose to use, remember that the most important consideration (after semantic and syntactic correctness of course) is readability. You don’t need to make your SQL query as compact as possible, or as short as possible. You need to make it as easy for someone else to read and understand as possible.

### Q2. List the last name, job title and department name for all employees who work in Toronto.

Looking at the schema we can see that the columns we need for this query are all in different tables. last\_name is in the employees table, job\_title is in the jobs table, department\_name in the departments table and the city in the locations table. We will therefore need to join all these table together.

The join type will be inner joins because we want to find the rows that match so we find the right job title for the employee and the city for their department etc. Finally, because the question asks us to limit out results to the employees working in Toronto, we know we will have a WHERE clause to filter the results at the end.

Putting it all together we get the following SQL:

SELECT last\_name, job\_title, e.department\_id, department\_name

FROM employees e

INNER JOIN departments d

ON e.department\_id = d.department\_id

INNER JOIN jobs j

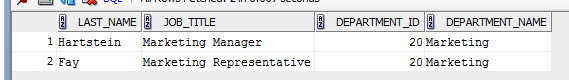
ON e.job\_id = j.job\_id

INNER JOIN locations l

ON d.location\_id = l.location\_id

WHERE l.city = ‘Toronto’;

The results should be:



Notice that we have used aliases for the table names. This is because the join conditions always use columns with the same name in both tables and so we must specify the table name in order to avoid ambiguity. To make the query more readable we can use sensible aliases so that we can then write e.job\_id rather than the full employees.job\_id.

An alternative way to write the same query would be to make use of the USING keyword:

SELECT last\_name, job\_title, department\_id, department\_name

FROM employees

INNER JOIN departments USING(department\_id)

INNER JOIN jobs USING (job\_id)

INNER JOIN locations USING (location\_id)

WHERE city = 'Toronto';

Personally, I think this is neater than using ON and is just as clear. However, it is rarer to see USING than ON because it can only be used when the join columns have the same name whereas ON can always be used. Some may argue also that ON is more verbose and therefore clearer and ON is supported by all relational databases whereas some may not support USING (although all the major ones do).

### Q3. Show all employees (their employee IDs and names) and their managers (their manager IDs and names). In both cases, combine first names and last names into one field.

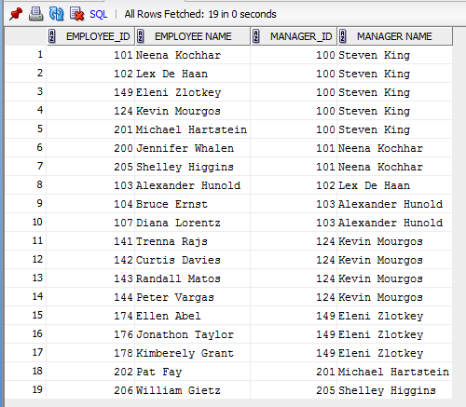
The tricky bit of this question is that the employees and the managers are both in the same table but on different rows. We know by now how to combine rows from different tables but how can we combine rows in the same table?

**NEW MATERIAL**

The answer is to pretend that there are two tables called employees and to join them together. When we write a query that joins a table with itself it is called a **self-join**. There is only one version of the table, obviously, but conceptually we can think of it as if there were two copies. We will have to use aliases for the table names because otherwise we will have two tables with the same name and all the columns will have the same names so everything will be ambiguous.

Once we understand that this is a self-join then the SQL is fairly standard:

The results should be:



Notice that we use e and m as table aliases to stand for employee and manager. This means that when we do the join condition it is obvious that we need the manager ID from e and the employee ID from m. If we had chosen less clear aliases, say e1 and e2, then that would have been harder to figure out. If we had put them the wrong way around then we would end up with a syntactically correct query but one that was semantically wrong and gave the wrong results.

### Q4. Management is considering a major restructuring. They are considering creating local versions of every department in every country the company operates in. List the new departments that would have to be created in each country.

This question is quite complex so it’s helpful to think about what the output should look like. If we run “SELECT \* FROM departments;” we will see that currently there is only a single version of each department in a single location. If we ran a query which joined departments, locations and countries we could list the departments and the names of the countries they are located in. The SQL would be:

SELECT department\_name, country\_name

FROM departments

NATURAL JOIN locations

NATURAL JOIN countries;

We can use the natural join because this query is only for our own understanding, not for production code.

The result of this query is:



The question is asking for a list of new departments to be created in each country. If we run “SELECT country\_name FROM countries;” we will get a list of the four countries that the company operates in (USA, UK, Canada and Germany). So if the company currently has an Administration department in the USA, the question is asking us to list that it needs to create an Administration department in Canada, Germany and the UK. What we are being asked to do is combine every possible combination of department\_name and country\_name and then remove those combinations that already exist in the database.

First, let us consider how to generate every possible combination. In fact, this is a straightforward application of the Cross Join:

SELECT department\_name, country\_name

FROM departments

CROSS JOIN countries;

Next, how will we remove the combinations that already exist? Since we are talking about removing rows from the result we need a WHERE clause. We want to remove any row where the combination of department\_name and country\_name already appears. Unfortunately, there is no common column between departments and countries and so we cannot usefully perform an inner join between these two tables to find out which countries each department is in at the moment. We will have to join the locations table as well. This will allow us to list, for each department in the departments table, its country\_id which we can get from the locations table. If we do this then we can use the WHERE clause to remove rows where the country\_id of the department matches the country\_id of the country we are generating a new pair with via the Cross Join.

But we don’t want to actually show the location\_id or country\_id. Fortunately, we can include a join without actually displaying any columns from the joined table. To make the results more readable we can also order them alphabetically by the name of the department. Our SQL is therefore:

SELECT department\_name, country\_name

FROM departments

CROSS JOIN countries

INNER JOIN locations USING (location\_id)

WHERE locations.country\_id != countries.country\_id

ORDER BY department\_name;

The results should be:



### Q5. Show all employees (their employee IDs and names) and their managers (their manager IDs and names). In both cases, combine first names and last names into one field. If an employee has no manager, still include the employee.

This is essentially the same as Q3 above but this time we have to include all employees, even those without managers. Recall our answer to Q3:

SELECT e.EMPLOYEE\_ID,

e.FIRST\_NAME || ' ' || e.LAST\_NAME AS "EMPLOYEE NAME",

m.EMPLOYEE\_ID AS "MANAGER\_ID",

m.FIRST\_NAME || ' ' || m.LAST\_NAME "MANAGER NAME"

FROM EMPLOYEES e

INNER JOIN EMPLOYEES m

ON e.MANAGER\_ID = m.EMPLOYEE\_ID;

For that question, we used an inner join to perform the self-join. The inner join will only return rows that match based on the join condition. But if an employee has no manager then the manager\_id on their row will be NULL and NULL won’t match, so that row will not be returned at all.

The solution is to use the left join which will return the row from the left table even if there is no match in the right table. Recall that in SQL the left table is the one we list first. So all we need to do is use the same SQL as for Q3 but change INNER to LEFT:

SELECT e.EMPLOYEE\_ID,

e.FIRST\_NAME || ' ' || e.LAST\_NAME AS "EMPLOYEE NAME",

m.EMPLOYEE\_ID AS "MANAGER\_ID",

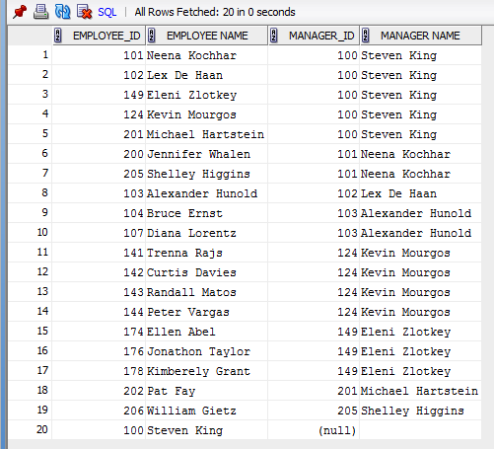
m.FIRST\_NAME || ' ' || m.LAST\_NAME "MANAGER NAME"

FROM EMPLOYEES e

**LEFT OUTER** JOIN EMPLOYEES m

ON e.MANAGER\_ID = m.EMPLOYEE\_ID;

The result is now:



# Completion Problems

### Q1. List the job title and minimum salary for every job where the minimum salary is more than 10,000. Combine the output with every country the company operates in.

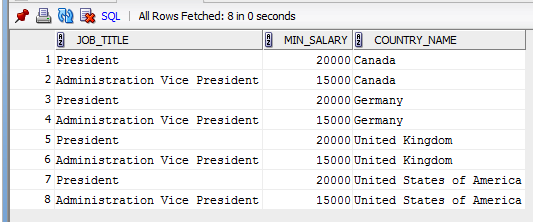
A: SELECT job\_title, min\_salary, country\_name   
FROM jobs, countries  
WHERE min\_salary > 10000;

SELECT JOB\_TITLE, MIN\_SALARY, COUNTRY\_NAME

FROM JOBS

CROSS JOIN COUNTRIES

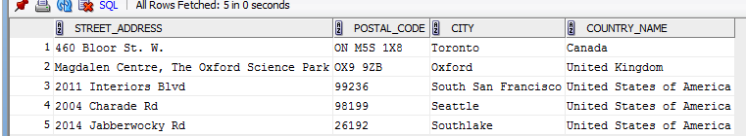
WHERE MIN\_SALARY > 10000;



### Q2. List the full address of every location, including the country name.

A: SELECT street\_address, postal\_code, city, country\_name   
FROM locations

INNER JOIN COUNTRIES

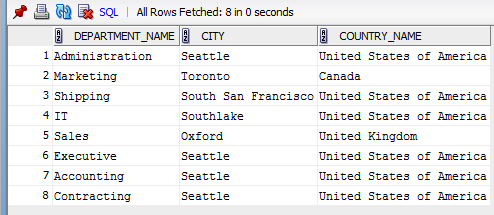
ON LOCATIONS.COUNTRY\_ID = COUNTRIES.COUNTRY\_ID;  


### Q3. List the name, city and country of every department.

A: SELECT department\_name, city, country\_name   
FROM departments

INNER JOIN LOCATIONS

USING (LOCATION\_ID)  
INNER JOIN countries   
 USING (country\_id);



### Q4. List the department ID, name and manager name for every department.

A: SELECT d.department\_id, department\_name,

first\_name || ' ' || last\_name AS Manager

FROM departments d

INNER JOIN employees e

ON d.MANAGER\_ID = e.EMPLOYEE\_ID;

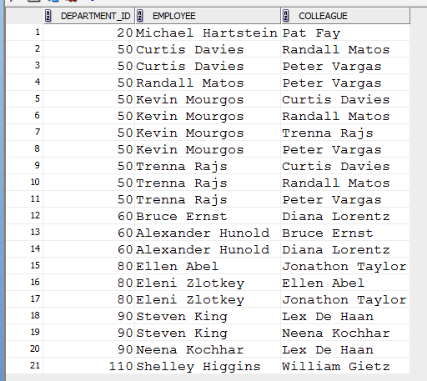
### 

### Q5. List every pair of employees working in the same department, but only list each pair once. Order the results by department ID then by the surname of the first employee and then the surname of the second employee.

A: SELECT e.department\_id, e.first\_name || ' ' || e.last\_name Employee,   
 c.first\_name || ' ' || c.last\_name Colleague   
FROM employees e   
INNER JOIN employees c

ON e.department\_id = c.department\_id

WHERE e.employee\_id < c.employee\_id  
ORDER BY e.department\_id, e.last\_name, c.last\_name;



### Q6. List every employee, their department name and manager name. If they have no manager or department, still include them.

A: SELECT e.first\_name || ' ' || e.last\_name AS "employee name", department\_name,

m.first\_name || ' ' || m.last\_name AS "manager name"

FROM employees e

LEFT OUTER JOIN departments d

ON e.department\_id = d.department\_id

LEFT JOIN EMPLOYEES m

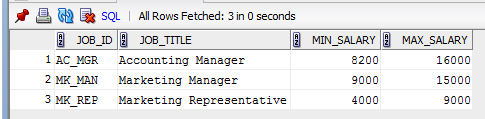
ON m.employee\_id = e.manager\_id

ORDER BY e.last\_name;



### Q7. List the details for all jobs with the letters “ing” in the title.

A: SELECT \*   
FROM jobs

WHERE job\_title LIKE ‘%ing%’;  


### Q8. List the information about employees who have left the company (i.e. have an end date).

A: SELECT first\_name, last\_name, start\_date, end\_date

FROM employees  
INNER JOIN job\_history   
 USING (employee\_id);

### 

# Deliberate Practice: Write the SQL

### Q1. List the names of all employees and their job titles, ordered alphabetically by surname.

SELECT first\_name, last\_name, job\_title

FROM employees e

INNER JOIN jobs j

ON e.job\_id = j.job\_id

ORDER BY last\_name ASC;

### 

### Q2. List the names of the three employees with the highest potential salary and their job title

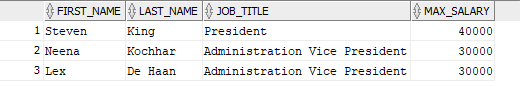
SELECT first\_name, last\_name, job\_title, max\_salary

FROM employees

INNER JOIN jobs

USING(job\_id)

FETCH FIRST 3 ROWS ONLY;



### Q3. The HR department wants a list of all employees who were hired after their managers. Produce a query that shows the names and hire dates of managers and the employees who they manage. Only show employees who were hired after their managers.

SELECT m.FIRST\_NAME || ' ' || m.LAST\_NAME "MANAGER NAME", m.hire\_date AS "Hire date", e.FIRST\_NAME || ' ' || e.LAST\_NAME AS "EMPLOYEE NAME",

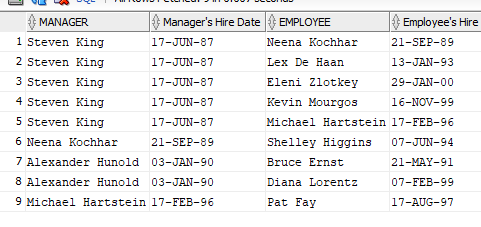
e.hire\_date AS "Hire date"

FROM EMPLOYEES e

INNER JOIN EMPLOYEES m

ON e.MANAGER\_ID = m.EMPLOYEE\_ID

WHERE m.hire\_date < e.hire\_date;;



### Q4. List the job titles and names of employees who used to hold those positions, including their start and end dates. Where nobody has ever held that position only list the job titles and where the position is still being filled list the job title, employee name and start date. Give the results in order of job title.

SELECT j.job\_title,e.first\_name, e.last\_name, h.start\_date, h.end\_date

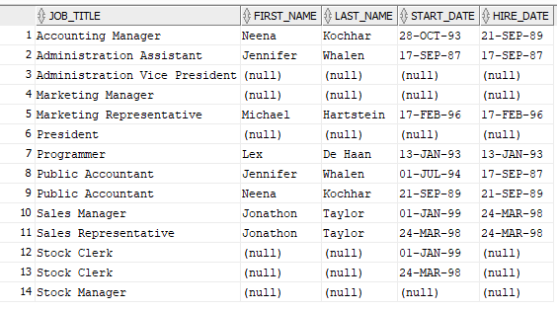
FROM jobs j

LEFT JOIN job\_history h

ON j.job\_id = h.job\_id

LEFT JOIN employees e

ON h.employee\_id = e.employee\_id

ORDER BY job\_title ASC;

### Q5. List the names of all employees still working for the company.

SELECT first\_name, last\_name

FROM employees

LEFT JOIN job\_history USING (employee\_id)

WHERE end\_date IS NULL;

### 

### Q6. List all employees who left the company in 1999

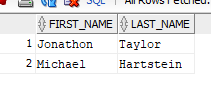
SELECT e.first\_name, e.last\_name

FROM employees e

INNER JOIN job\_history h

ON e.employee\_id = h.employee\_id

WHERE EXTRACT(YEAR FROM h.end\_date) = 1999;



### Q7. List the manager name, department name and full address including country and region name for every department – including those with no manager.

SELECT m.first\_name|| ' ' ||m.last\_name AS Manager, d.department\_name, l.street\_address, postal\_code, city, state\_province, c.country\_name, r.region\_name

FROM departments d

LEFT JOIN employees m

ON m.department\_id = d.department\_id

LEFT JOIN employees e

ON m.manager\_id = e.employee\_id

LEFT JOIN locations l

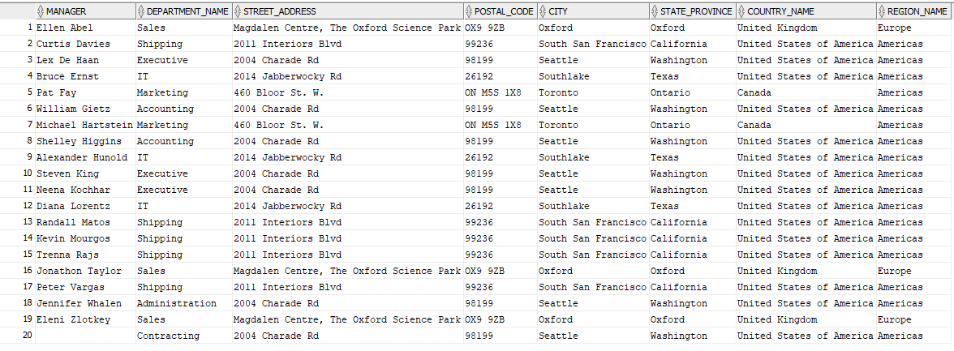
ON d.location\_id = l.location\_id

LEFT JOIN countries c

ON l.country\_id = c.country\_id

LEFT JOIN regions r

ON c.region\_id = r.region\_id;



### Q8. List the employees who are in the upper half of their job’s salary range, include their job title, their salary and the mid-point of their job’s salary range.

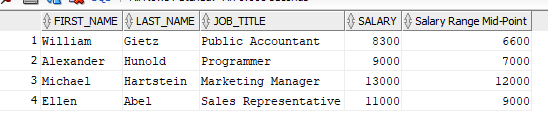
SELECT first\_name, last\_name, job\_title, salary,(min\_salary + max\_salary)/2

FROM employees e

INNER JOIN jobs j

ON e.job\_id = j.job\_id

WHERE salary > (min\_salary + max\_salary)/2;



### Q9. List the names of all employees and the countries where they work.d

SELECT first\_name, last\_name, country\_name

FROM employees e

INNER JOIN departments d

ON e.department\_id = d.department\_id

INNER JOIN locations l

ON d.location\_id = l.location\_id

INNER JOIN countries c

ON l.country\_id = c.country\_id;

