Group 3:

**SQL-Mongo Project –**

**IBM HR Analytics Employee Attrition & Performance**

BUAN 6320

Group Members

Naga Jyothi Muvva

Lasya Bodapati

Manasseh Jesrun Kota

Venkata Ganga Govardhan Mowva

Praneeth Chandra Budala

Pranay Kumar Reddy Yammanuru

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Jyothi** | **Lasya** | **Jesrun** | **Govardhan** | **Praneeth** | **Pranay** |
| Prepared Data Model and Created Physical DB | x | x |  | x | x |  |
| Loaded Data into Database | x |  | x | x |  | x |
| Wrote SQL Queries |  | x | x | x | x | x |
| Prepared Mongo Database |  | x |  |  | x | x |
| Loaded data into Mongo DB | x |  | x |  |  | x |
| Wrote Mongo Queries | x |  | x | x | x |  |
| Prepared Report |  | x | x | x |  | x |
| Reviewed Report | x | x |  |  | x |  |

Contents

Relational Data Model 4

Assumptions/Notes About Data Entities and Relationships 4

Entity-Relationship Diagram 4

Physical MySQL Database 5

Assumptions/Notes About Data Set 5

Screen shot of Physical Database objects 5

Data in the Database 5

SQL Queries 6

SQL Query 1 6

Question 6

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 6

Translation 6

Screen Shot of SQL Query and Results 6

SQL Query 2 7

Question 7

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 7

Translation 7

Screen Shot of SQL Query and Results 7

SQL Query 3 8

Question 8

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 8

Translation 8

Screen Shot of SQL Query and Results 8

SQL Query 4 9

Question 9

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 9

Translation 9

Screen Shot of SQL Query and Results 9

SQL Query 5 10

Question 10

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 10

Translation 10

Screen Shot of SQL Query and Results 10

Data Review for MongoDB 11

Assumptions/Notes About Data Collections, Attributes and Relationships between Collections 11

Physical Mongo Database 12

Assumptions/Notes About Data Set 12

Screen shot of Physical Database objects (Database, Collections and Attributes) 12

Data in the Database 12

MongoDB Queries/Code 13

Mongo Query 1 13

Question 13

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 13

Translation 13

Screen Shot of MongoDB Query/Code and Results 13

Mongo Query 2 14

Question 14

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 14

Translation 14

Screen Shot of MongoDB Query/Code and Results 14

Mongo Query 3 15

Question 15

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 15

Translation 15

Screen Shot of MongoDB Query/Code and Results 15

# **Relational Data Model**

# The following are the attributes removed from the data:

* **EmployeeCount: This field was removed as there are constant values throughout the employees (in this case “1”), this signifies redundant data.**
* **Over18: This field was removed from the data as the user can retrieve above 18 age groups from the “Age” field.**
* **StandardHours: This field was removed as there are constant values throughout the employees (in this case “80”), this also signifies redundancy. Also, this field lacks statistical significance.**

# **Entities Description:**

* **Employee:** This table consists of the entire information of all the employees. In this entity, the “EmployeeNumber” is set as a primary key. In other words, this table fulfils 3NF criteria since all other attributes are dependent on the one primary key.
* **EducationField: This table consists of different educational backgrounds of all the employees. For this entity, the “**EducationFieldID” is set as a primary key. There are only two attributes in the entity, the other attribute is dependent on primary key therefore the table exists in 3NF.
* **EducationDegree: This table consists of different educational levels of all the employees (Bachelors, Masters, Below college, college etc.). Here, “EducationDegreeID” is set as a primary key. All the employees share an associative relationship with the primary key; therefore, the entity exists as 3NF.**
* **EducationDetails: This table has many to many relationships with the entities – EducationField and EducationDegree. EducationDetails shares an associative relationship with the other two entities, which exists as 3NF.**
* **JobRoles: This table consists of all the job roles pertaining to the employees. Here, the JobRoleID is set as a primary key. The table consists of only two attributes, where there is a dependency on the primary key. Also, this entity is created from the attribute** “JobRole”. Therefore, this exists as 3NF.
* **Department: This table consists of different departments based on the employees’ job responsibilities. Here, the “DepartmentID” is set as a primary key. The table is derived from the attribute “Department” from the given dataset. There are only 2 attributes in ths table with an associative relationship. Therefore, the data exists as 3NF.**
* **WorkExperience: This table consists of work experience of all the employees who are working within and outside the organization. Here, the “EmployeeNumber” is set as a primary key for the WorkExperience entity. It also exists as a foreign key. The attributes are dependent on the work experience of the employees which is the primary key, therefore the table exists as 3NF.**
* **Salary: This table consists of the pay rates of the employees (hourly,daily). Here, “SalaryID” is set as a primary key. All the attributes are dependent on the primary key of the table; therefore this entity exists as a 3NF.**
* **Feedback: This entity consists of different commitments and promises made by the employees (attrition, organizational behavior) to the company. Here, “EmployeeNumber” is set as a primary key with all the other attributes depending on it. Therefore the entity exists as a 3NF.**
* **WorklifeBalance: This entity consists of different levels of worklife balance from a scale of 1 to 4 which describes from bad to good. Here, “WorkLifeBalance” is set as a primary key where all the attributes are dependent on. Therefore, this entity exists as a 3NF.**
* **PerformanceRating: This entity consists of different performance ratings on a 4-level scale (low, good, excellent, outstanding). Here, “PerformanceRating” is set as a primary key and the other attribute is dependent on the primary key. The table exists as a 3NF.**
* **JobInvolement: This entity consists of different levels at which the employee is involved in the job aspect (low, medium, high, very high). Here, “JobInvolevementLevel” is set as a primary key where the other attribute is dependent on, therefore the table exists as a 3NF.**
* **JobSatisfaction: This entity consists of different levels of job satisfaction from the employees point of view (low, medium, high, very high), here the “JobInvolvementLevel” is set as a primary key. The other attribute is dependent on the primary key; therefore, the table exists as a 3NF.**

# **Relationships between entities:**

|  |  |  |
| --- | --- | --- |
| **ENTITIES** | **KEYS** | **TYPE** |
| Employee | EmployeeNumber | Primary key |
| JobRoleID | Foreign key |
| DepartmentID | Foreign key |
| EducationField | EducationFieldID | Primary key |
| EducationDegree | EducationDegreeID | Primary key |
| JobRoles | JobRoleID | Primary key |
| Department | DepartmentID | Primary key |
| WorkExperience | EmployeeNumber | Primary key |
| Salary | SalaryID | Primary key |
| EmployeeNumber | Foreign key |
| Feedback | EmployeeNumber | Primary key |
| WorkLifeBalance | Foreign key |
| JobSatisfaction | Foreign key |
| PerformanceRating | Foreign key |
| JobInvolvementLevel | Foreign key |
| WorkLifeBalance | WorkLifeBalance | Primary key |
| PerformaceRating | PerformanceRating | Primary key |
| JobInvolvement | JobInvolvementLevel | Primary key |
| JobSatisfaction | JobSatisfaction | Primary key |

# Linking Tables Used:

|  |  |  |
| --- | --- | --- |
| **Linking Table** | **Keys** | **Type** |
| EducationDetails | EducationDegreeID EducationFieldID  EmployeeNumber | Composite Primary Key |

# Relationships:

|  |  |
| --- | --- |
| Employee to JobRoles | Many to one (n:1) |
| Employee to EducationDetails | One to one (1 :1) |
| Employee to Department | Many to one (n:1) |
| Feedback to Employee | One to One (1:1) |
| Salary to Employee | One to One (1:1) |
| WorkExperience to Employee | One to one (1:1) |
| EducationField to EducationDegree | Many to many (n:m) |
| WorkLifeBalance to Feedback | One to many (1: n) |
| Feedback to PerformanceRating | Many to one (n:1) |
| Feedback to JobInvolvement | Many to one (n:1) |
| Feedback to JobSatisfaction | Many to one (n:1) |

# Data Model

# A screenshot of a computer Description automatically generated

# Physical MySQL Database

## Assumptions/Notes About Data Set

* **Multiple “Education Degree” have multiple “Education fields”, there are many to many relationships between them. So, it is linked with the associate table Education Details.**
* **Each “job level” has multiple “job roles”.**
* **“Over18” can be retrieved from “Age” attribute, therefore it has been dropped as it has redundant data.**
* **Since Working Hours are not provided for each employee other than Standard Hours, Hourly Rate, Overtime, Daily Rate, Monthly Rate and MonthlyIncome. These can’t be derived from each other because of missing some data like Working hours of employee and Overtime Rate. These are ambiguous attributes and require more information to define them.**
* **There are no empty fields in the data set.**
* **Over18, EmployeeCount and StandardHours fields have fixed values hence it is redundant data then these fields are dropped.**
* StockOptionLevel means “how much company stocks each employee owns” since the dataset has the values 0,1,2,3 which are not clearly interpreted. Since most of the employees have 0 then this field has sparse data.
* When Attrition = 'Yes' for an employee, the record provides data about the employee's state at the time of leaving: Monthly income, Life Balance, Job Satisfaction etc., So these fields are mentioned in Feedback table.
* Relationship satisfaction means “how happy is the employee with her colleagues” Since this field is related to feedback then it is mentioned in Feedback table.
* “NumCompaniesWorked” field means how many companies have been worked by employee since it is related experience of employee then it is mentioned in Work Experience Entity.
* “EmployeeCount” field has fixed value for each value which is bad data, so it is dropped from the dataset.
* “PercentSalaryHike” field is the % change in salary from 2016 vs 2015, since it is related to salary Hike of each employee then it is placed in Salary Entity.

## Screen shot of Physical Database objects

**Employee**:

A screenshot of a computer

Description automatically generated

**EducationDegree**:

A screenshot of a computer

Description automatically generated

**EducationField**:

A screenshot of a computer

Description automatically generated

**EducationDetails**:

A screenshot of a computer

Description automatically generated

**JobRoles**:

A screenshot of a computer

Description automatically generated

**Department**:

A screenshot of a computer

Description automatically generated

**WorkExperience**:

A screenshot of a computer

Description automatically generated

**Salary**:

A screenshot of a computer

Description automatically generated

**Feedback**:

A screenshot of a computer

Description automatically generated

**JobSatisfaction**:

A screenshot of a computer

Description automatically generated

**WorkLifeBalance**:

A screenshot of a computer

Description automatically generated

**PerformanceRating**:

A screenshot of a computer

Description automatically generated

**JobInvolvement**:

A screenshot of a computer

Description automatically generated

## Data in the Database

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| Employee | EmployeeNumber | JobRoleID DepartmentID | 1470 |
| EducationDegree | EducationDegreeID |  | 5 |
| EducationField | EducationFieldID |  | 6 |
| EducationDetails | EducationDegreeID EducationFieldID  EmployeeNumber | EducationDegreeID EducationFieldID  EmployeeNumber | 1470 |
| JobRoles | JobRoleID |  | 9 |
| Department | DepartmentID |  | 3 |
| WorkExperience | EmployeeNumber | EmployeeNumber | 1470 |
| Salary | SalaryID | EmployeeNumber | 1470 |
| Feedback | EmployeeNumber | EmployeeNumber  WorkLifeBalance  JobSatisfaction  PerformanceRating  JobInvolvementLevel | 1470 |
| WorkLifeBalance | WorkLifeBalance |  | 4 |
| PerformaceRating | PerformaceRating |  | 4 |
| JobInvolvement | JobInvolvementLevel |  | 4 |
| JobSatisfaction | JobSatisfaction |  | 4 |

# 

# SQL Queries

**Pick 5 out of the 12 statements to write your queries in MySQL.**

## SQL Query 1

### Question

Which department's employee is the most likely to have the shortest commute between home and work?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

In this query we computed the average distance from home for each department and found out that Human Resource Department has less average distance. Therefore, this is the most likely to have the shortest commute between home and work.

There is only 1 row displaying in the result which department has the shortest commute between home and work.

### Translation

Select the DepartmentID, DepartmentName, and the average distance from home for employees from the Employee table, inner joining it with the WorkExperience table on the EmployeeNumber. Furtherly, join the result with the Department table on the DepartmentID. Group the results by DepartmentID and order them by the average distance. Finally, limit the obtained result to the first row.

### Screen Shot of SQL Query and Results

A screenshot of a computer

Description automatically generated

## SQL Query 2

### Question

A new employee from a Medical-related education field wants to work in Sales. Do you believe the company might be able to give her a chance to work in Sales? Why or Why not?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

The Average Performance of employees in the Sales department with respect to their education fields was calculated below in which medical related stands in the 2nd position whereas the 1st position is not for a specific field. Therefore, we can expect the company might give them a chance to work in Sales.

There are 5 rows displayed in the result for sales department by Education field with Average performance rating.

### Translation

Select the Department ID, Department Name, Education Field, and the average performance rating for employees from the Department table, joining it with the Employee table on the DepartmentID. Then, join the result with the Feedback table on the EmployeeNumber. Join the Education details table on the EmployeeNumber. Finally, join the EducationField table on the EducationFieldID. Filter the results to include only the department name with the description “Sales” and “Female” employees using where clause. Group the results by DepartmentID, DepartmentName, and EducationField, and order them by the average performance rating in descending order.

### Screen Shot of SQL Query and Results

A screenshot of a computer

Description automatically generated

## SQL Query 3

### Question

The HR department feels they have the highest job satisfaction while Research & Development department feels their department has the highest environment satisfaction. Who is right?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

From the below results, HR department is not correct because its Job satisfaction is lower compared to other departments and Research & Development department is right since its Environment Satisfaction is high.

There are 3 rows displayed in the result.

### Translation

Select the Department ID, Department Name, average job satisfaction, and average environment satisfaction for employees from the Department table, joining it with the Employee table on the DepartmentID. Then, join the result with the Feedback table on the EmployeeNumber. Group the results by DepartmentID.

### Screen Shot of SQL Query and Results

A screenshot of a computer

Description automatically generated

## SQL Query 4

### Question

An employee in Sales department has complained to HR saying that females are paid less than males in the company, in all departments. What insight can you provide to prove or disprove that statement?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

As per the below query results, Females are the highest paid on an average in all the departments. Hence the above statement proved wrong.

There are 3 rows displayed in the result.

### Translation

Select all columns from the result of the subquery, which calculates the average salary for each department and gender combination and assigns a rank to each combination based on the average salary. The subquery joins the Employee and Department tables on DepartmentID, and further joins Salary table on EmployeeNumber keys and group by department Id, Department Name and Gender. Finally, filter the result to only include rows where the maximum rank is 1.

### Screen Shot of SQL Query and Results

A screenshot of a computer

Description automatically generated

## SQL Query 5

### Question

A press article in a business magazine has said that at this company, married men have higher performance ratings than divorced or single men. What initial finding can you obtain from the data to help articulate the company's response in this regard?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

By computing the average performance rating for Male, from the results, we can say that married men performance rating is lower than Single men and higher than divorced men. Hence, the company can respond that the claim is false.

There are 3 rows displayed in the result.

### Translation

Select the Gender, MaritalStatus, and average performance rating for male employees from the Employee table, joining it with the Feedback table on the EmployeeNumber. Filter the results to include only male employees using where clause. Group the results by MaritalStatus and order them by the average performance rating in descending order.

### Screen Shot of SQL Query and Results

## A screenshot of a computer Description automatically generated

# Data Review for MongoDB

## Assumptions/Notes About Data Collections, Attributes and Relationships between Collections

**Data Collection**:

MongoDB documents tend to have all data for a given record in a single document, whereas in a relational database information for a given record is usually spread across many tables. The complete document can be accessed with a single call to the database, rather than having to JOIN multiple tables to respond to a query. In my db, the only collection with the name Employee\_Data.

In SQL database, each entity has different tables. But in MongoDB, each employee data is in one collection. However, in MongoDB, documents can contain embedded subdocuments, providing a much closer inherent data model to your applications. In my database, created only one collection which has all the employee data.

**Attributes**: The attributes Over18 and Employee Count are removed as it has duplicate data.

* Over 18 can be retrieved from the field age using aggregate function $gte:18
* Employee count has only one employee count data, therefore it is removed.

**Relationships**:

It has only one collection which has all employee documents with the name “Employee\_Data” Collection.

# Physical Mongo Database

## Assumptions/Notes About Data Set

* Documents are the basic unit of data storage in MongoDB.
* Each document is a JSON-like BSON object, where BSON is a binary representation of JSON-like documents.
* Over18 and Employee count is removed from the collection which has duplicate data.
* MongoDB supports various data types, including strings, integers, doubles, arrays, and nested documents in one collection.
* It is assumed that the collection has all necessary data about employees like personal information, Salary, travel, Role, Department, overall ratings, and Work Experience.
* MongoDB documents tend to have all data for a given record in a single document, whereas in a relational database information for a given record is usually spread across many tables.
* A single read to the database can retrieve the entire document containing all related data.
* BSON Documents make it simpler and faster for developers to model how data in the application will map to data stored in the database.
* It is assumed that dataset is trustworthy, and that data can be relied upon for making decisions.
* It is assumed that the data set can be handled with the proper degree of confidentiality and that any sensitive data like employee personal information will be safeguarded.
* It is assumed that dataset is consistent, which means that each field has set format and the data is constant across all documents.

## Screen shot of Physical Database objects (Database, Collections and Attributes)

A screenshot of a computer

Description automatically generated

## Data in the Database

|  |  |  |
| --- | --- | --- |
| **Collection Name** | **Relationships With Other Collections (if any)** | **# of Documents in Collection** |
| Employee\_Data |  | 1470 |

# MongoDB Queries/Code

Pick 3 SQL queries and write them in MongoDB

## Mongo Query 1

### Question

The company has been paying gas expenses for miles traveled by employees between their home and work. If they want to increase the per mile compensation, which department's employees will gain the most?

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

As per the below Mongo Query Result, found the employees average distance from home for each department. As per observation, Sales have higher miles. If they want to increase the per mile compensation, then the Sales department will gain the most compensation.

There are 3 documents displayed in the result for each department(Sales, Human Resources and Research & Development) with average distance from home.

### Translation

Using Employee\_Data collection, utilize aggregate function to group documents by Department (using Department field) and get an average distance from home.

### Screen Shot of MongoDB Query/Code and Results

**MongoDB Query:**

db.Employee\_Data.aggregate({$group: {\_id: "$Department", avg\_DistanceFromHome: {$avg: "$DistanceFromHome"}}})

A screenshot of a computer program

Description automatically generated

## Mongo Query 2

### Question

An employee from Medical education field working in Sales department has spread a rumor saying that employees with his educational background are paid more in Research & Development than in Sales. What insight can you provide to prove or disprove that statement?

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

As per below query result, After calculating average monthly salary for Medical field employees in Sales and Research & Development Departments, we can observe that average monthly income is more for Research & Development than in Sales. Hence the statement is proved.

There are 2 documents displayed in the result. One for Sales department, Medical field and another one for Research & Development, Medical Field.

### Translation

Using Employee\_Data collection, utilize aggregate function to match Education Field by Medical when Department in Sales, Research and Development and group documents by Department (using Department field) and Education Field and get an average Monthly Income.

### Screen Shot of MongoDB Query/Code and Results

**MongoDB Query:**

db.Employee\_Data.aggregate([{$match:{EducationField: "Medical", Department: {$in: ["Sales","Research & Development"]}}},{$group:{ \_id:{Department: "$Department", EducationField: "$EducationField"}, average\_Salary:{$avg:"$MonthlyIncome"}}}]);

A screenshot of a computer

Description automatically generated

## Mongo Query 3

### Question

A press article in a business magazine has said that at this company, single women in Sales have worked at the company longer than divorced or married women. What initial finding can you obtain from the data to help articulate the company's response in this regard?

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

As per the below query result, average number of years worked at the company for Single woman is less than the Married and Divorced women in sales department. Hence the statement has been proved wrong.

There are 3 documents displayed in the result. Gender Female for Marital Status Single, Married and Divorced for Sales Department.

### Translation

Using Employee\_Data collection, utilize aggregate function to match Gender Field by Female when Department in Sales and group documents by Gender field, Marital Status and Department Field and get an average YearsAtCompany.

### Screen Shot of MongoDB Query/Code and Results

**MongoDB Query:**

db.Employee\_Data.aggregate([{$match:{Gender:"Female", Department:{$in:["Sales"]}}},{$group:{\_id:{Gender:"$Gender",MaritalStatus:"$MaritalStatus",Department:"$Department"},average\_YearsAtCompany:{$avg:"$YearsAtCompany"}}}]);

A screenshot of a computer

Description automatically generated