Project -02
Employees Data

Submitted by Shridhar Ramani

Contents

1. Problem Statement
2. Objective
3. Why Big Data
4. Technology Stack
5. Steps
6. Results
7. Recommendations
8. Limitations

Employee Data

Problem Statement

The company is looking how its data can be transformed to get some actionable insights and rights set of data engineering technology stack. The database contains a huge volume of employees data with both employee engagement survey results and exit data. This provides a massive, globally diverse, and statistically relevant dataset for conducting research specific to attrition.

Objective –

To design an end-to-end data pipeline and analyze the data for the organization which is looking for updating its employees policies and dealing with the issues related to the attrition rates over the given time period.

Why it is a part of Bigdata -

Since the data volume is high and analyzing the data by using traditional approaches is time consuming .

Technology Stack

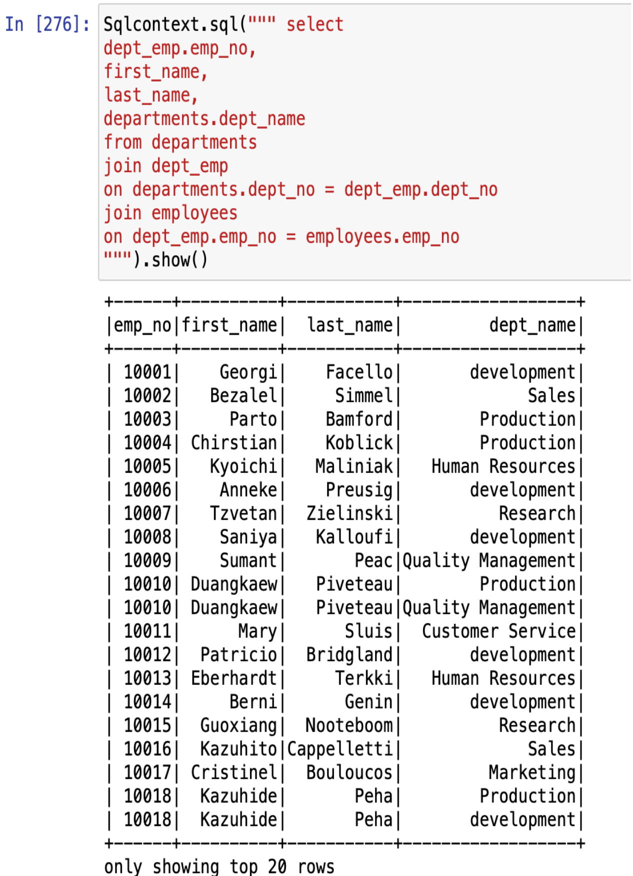
-MySQL (to create database) and table schemas

- Linux Commands
- Sqoop (Transfer data from MySQL Server to HDFS/Hive) - HDFS (to store the data)
- Hive (to create database)
- Impala (to perform the EDA)
- SparkSQL (to perform the EDA)
- SparkML (to perform model building)

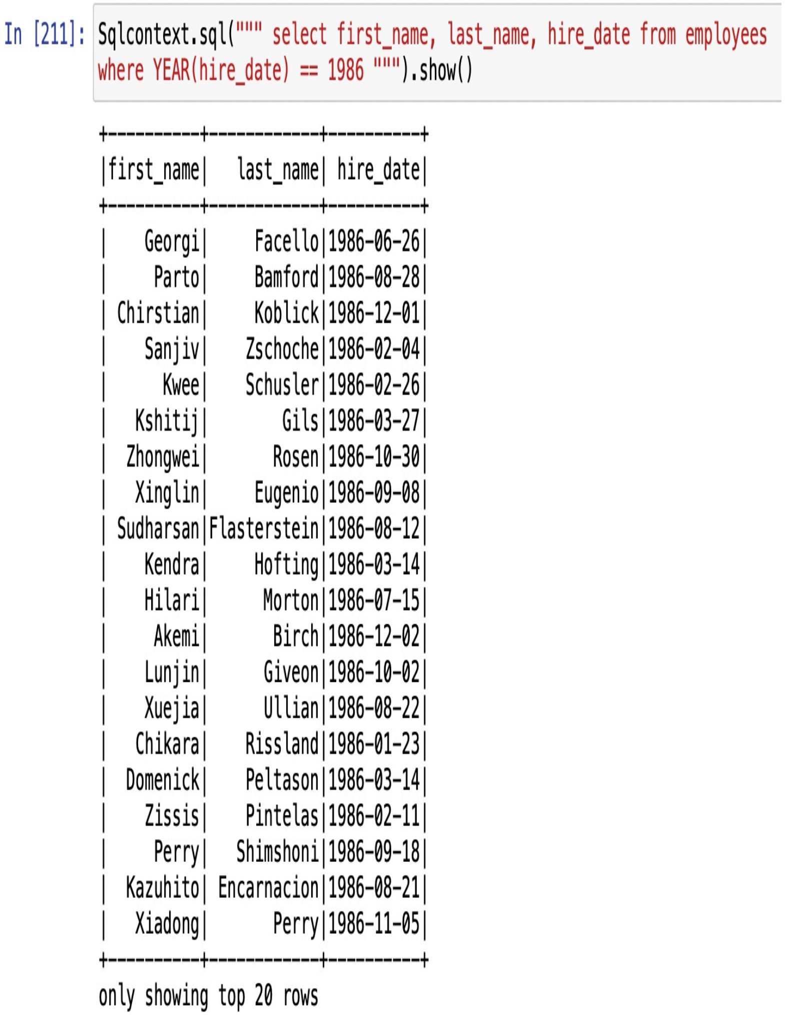
Steps

- Defining the business problem
- Weighing the tools and technology to use which is the most appropriate technology stack
- Mysql – Creating the tables on MySQL for the given data set which is CSV format. We create database ,schema and load the data on it.
- Go to the HDFS and delete any pre existing tables with the same name.
- Use the SQOOP and load on hdfs and data stored in paraquet file .
- In hive create a schema for all the tables and perform analysis
- Open Jupyter notebook and use Spark (sparksql and pyspark) for answering the business queries.
- Exploratory data analysis
- Create entire data pipeline

Results

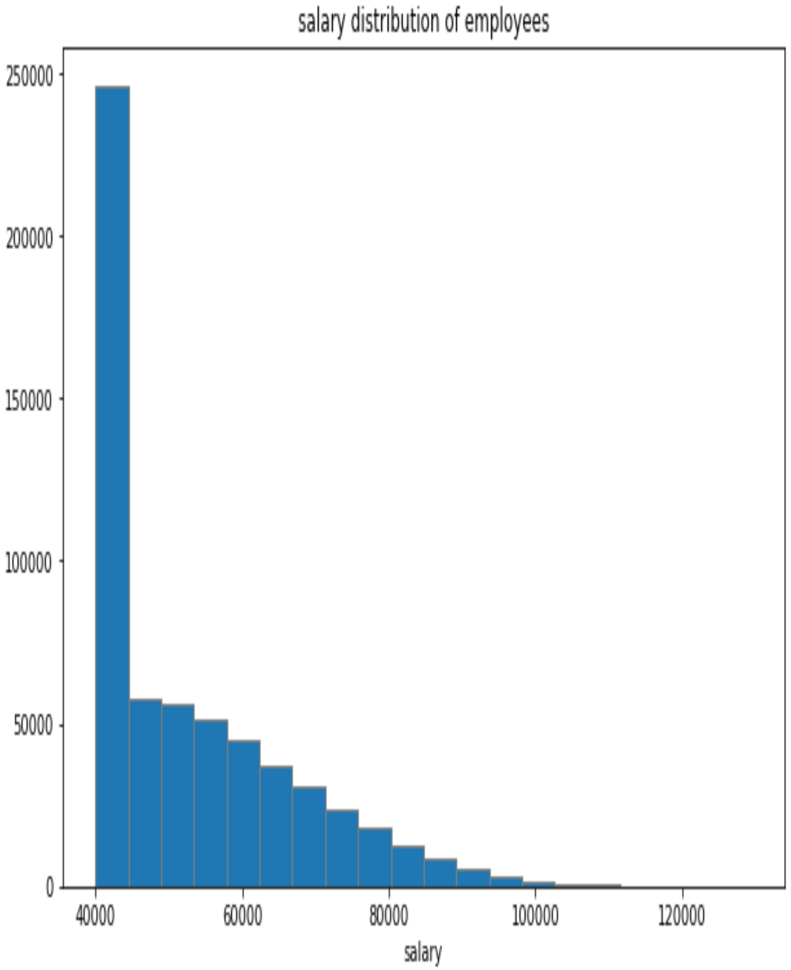


```
In [276]: Sqlcontext.sql(""" select
          dept_emp.emp_no,
          first_name,
          last_name,
          departments.dept_name
          from departments
          join dept_emp
          on departments.dept_no = dept_emp.dept_no
          join employees
          on dept_emp.emp_no = employees.emp_no
          """).show()

          +------+----------+-----------+------------------+
          |emp_no|first_name|  last_name|         dept_name|
          +------+----------+-----------+------------------+
          | 10001|    Georgi|    Facello|       development|
          | 10002|   Bezalel|     Simmel|             Sales|
          | 10003|     Parto|    Bamford|        Production|
          | 10004| Chirstian|    Koblick|        Production|
          | 10005|   Kyoichi|   Maliniak|   Human Resources|
          | 10006|    Anneke|    Preusig|       development|
          | 10007|   Tzvetan|  Zielinski|          Research|
          | 10008|    Saniya|   Kalloufi|       development|
          | 10009|    Sumant|       Peac|Quality Management|
          | 10010| Duangkaew|   Piveteau|        Production|
          | 10010| Duangkaew|   Piveteau|Quality Management|
          | 10011|      Mary|      Sluis|  Customer Service|
          | 10012|  Patricio|   Bridgland|       development|
          | 10013| Eberhardt|     Terkki|   Human Resources|
          | 10014|     Berni|      Genin|       development|
          | 10015|  Guoxiang|  Nooteboom|          Research|
          | 10016|  Kazuhito|Cappelletti|             Sales|
          | 10017|  Cristinel|  Bouloucos|         Marketing|
          | 10018|  Kazuhide|       Peha|        Production|
          | 10018|  Kazuhide|       Peha|       development|
          +------+----------+-----------+------------------+
          only showing top 20 rows
```

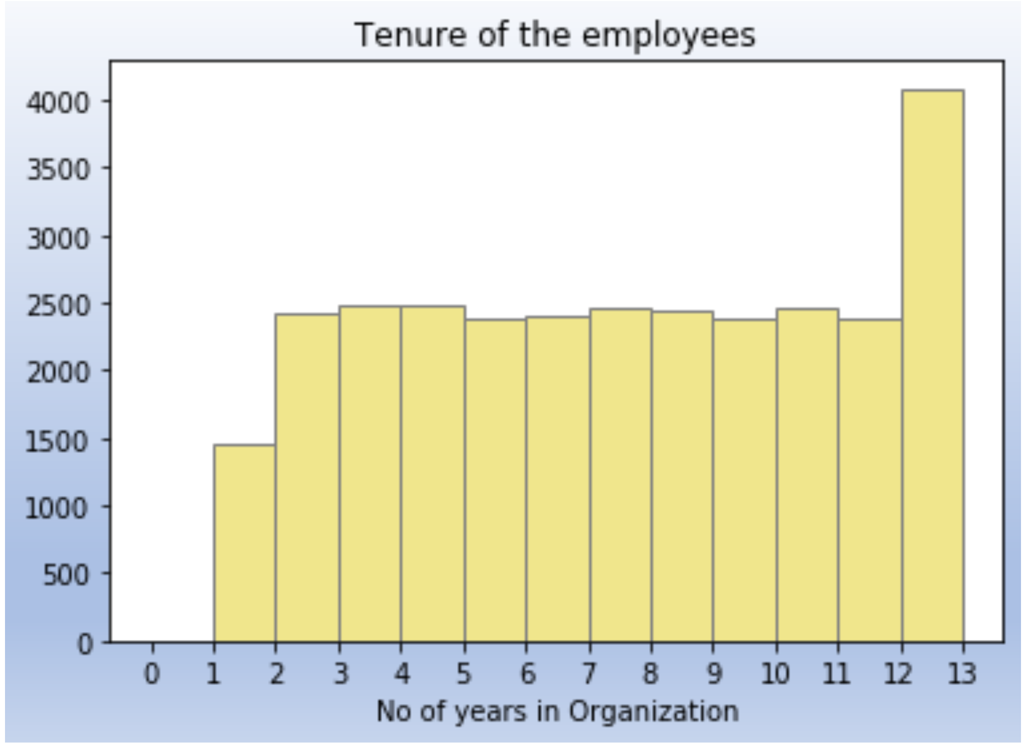


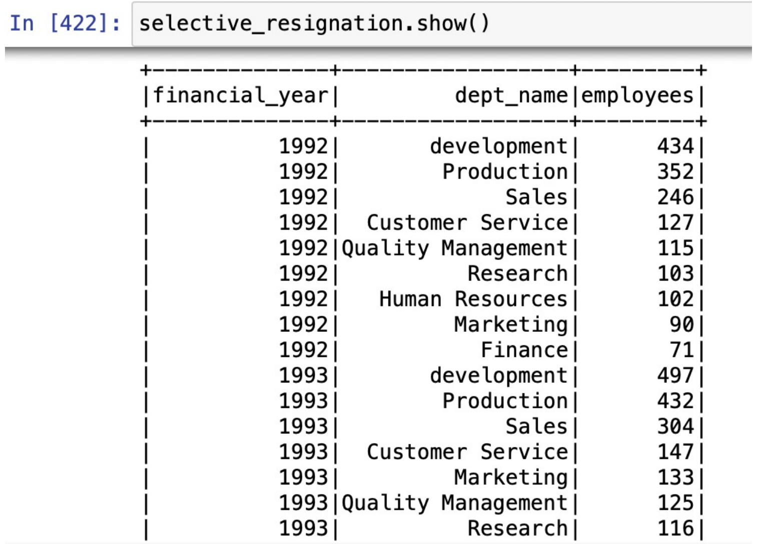
```
In [211]: Sqlcontext.sql(""" select first_name, last_name, hire_date from employees
          where YEAR(hire_date) == 1986 """).show()

          +----------+------------+----------+
          |first_name|   last_name| hire_date|
          +----------+------------+----------+
          |    Georgi|     Facello|1986-06-26|
          |     Parto|     Bamford|1986-08-28|
          | Chirstian|     Koblick|1986-12-01|
          |    Sanjiv|    Zschoche|1986-02-04|
          |      Kwee|    Schusler|1986-02-26|
          |   Kshitij|        Gils|1986-03-27|
          |  Zhongwei|       Rosen|1986-10-30|
          |   Xinglin|     Eugenio|1986-09-08|
          | Sudharsan|Flasterstein|1986-08-12|
          |    Kendra|     Hofting|1986-03-14|
          |    Hilari|      Morton|1986-07-15|
          |     Akemi|       Birch|1986-12-02|
          |    Lunjin|      Giveon|1986-10-02|
          |    Xuejia|      Ullian|1986-08-22|
          |    Chikara|    Rissland|1986-01-23|
          |  Domenick|    Peltason|1986-03-14|
          |    Zissis|    Pintelas|1986-02-11|
          |     Perry|   Shimshoni|1986-09-18|
          |  Kazuhito| Encarnacion|1986-08-21|
          |   Xiadong|       Perry|1986-11-05|
          +----------+------------+----------+
          only showing top 20 rows
```

Most of the employees salary is about 40k



About 4000 employees are working in the company for almost 13 years.
About 1500 employees are working for about 1 to 2 years .

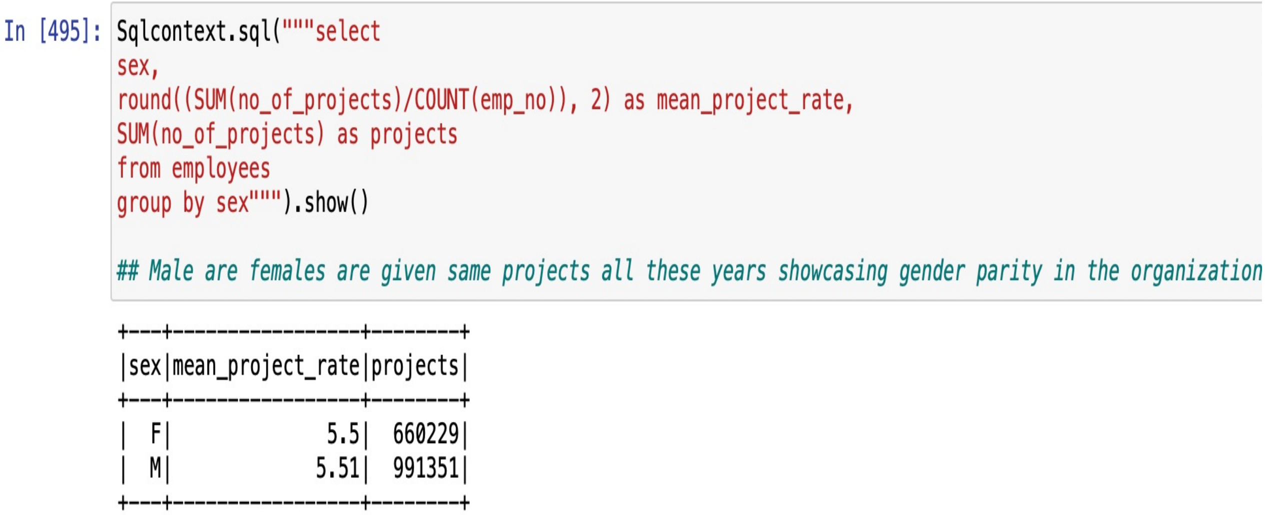




```
In [422]: selective_resignation.show()

+--------------+------------------+---------+
|financial_year|         dept_name|employees|
+--------------+------------------+---------+
|          1992|       development|      434|
|          1992|        Production|      352|
|          1992|             Sales|      246|
|          1992|  Customer Service|      127|
|          1992|Quality Management|      115|
|          1992|          Research|      103|
|          1992|   Human Resources|      102|
|          1992|         Marketing|       90|
|          1992|           Finance|       71|
|          1993|       development|      497|
|          1993|        Production|      432|
|          1993|             Sales|      304|
|          1993|  Customer Service|      147|
|          1993|         Marketing|      133|
|          1993|Quality Management|      125|
|          1993|          Research|      116|
```

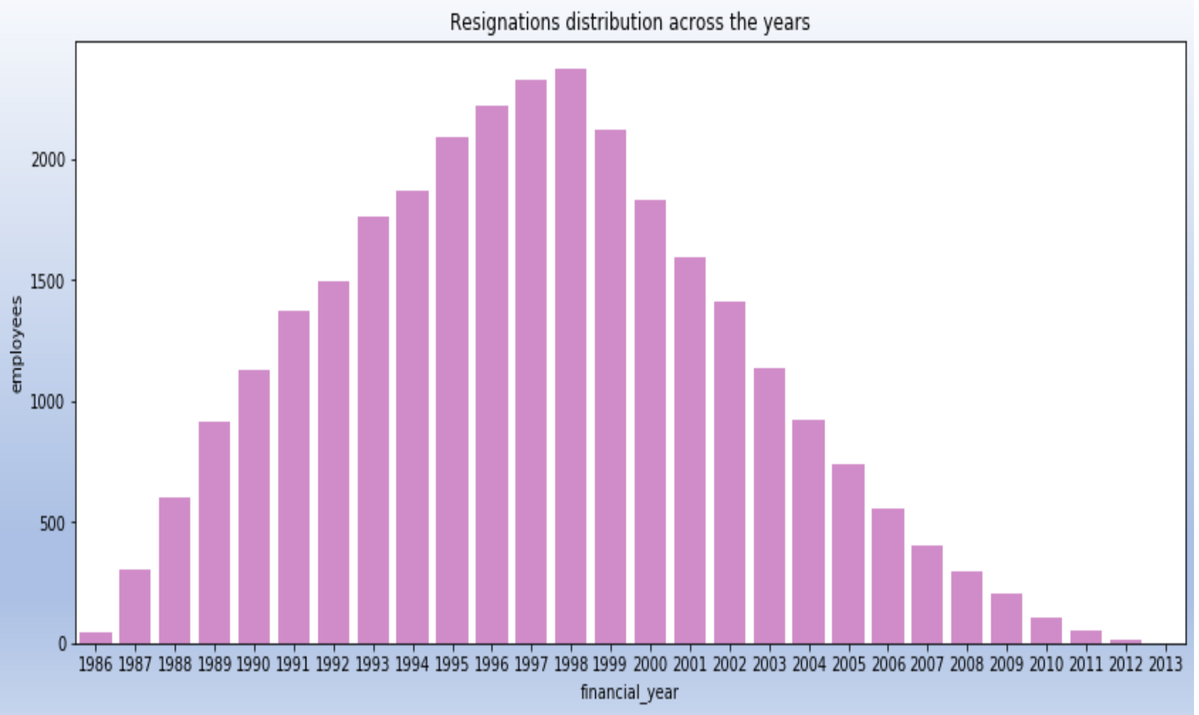
How many employees from respective departments
 resigned the most
About 497 employees resigned from development department hence it suggest some identification of issues and amendments in the department.

The mean project rate is almost equal for
Male and female employees



```
In [495]: Sqlcontext.sql("""select
          sex,
          round((SUM(no_of_projects)/COUNT(emp_no)), 2) as mean_project_rate,
          SUM(no_of_projects) as projects
          from employees
          group by sex""").show()

          ## Male are females are given same projects all these years showcasing gender parity in the organization
```

```
+---+-----------------+--------+
|sex|mean_project_rate|projects|
+---+-----------------+--------+
|  F|              5.5|  660229|
|  M|             5.51|  991351|
+---+-----------------+--------+
```



The employees resigned more and more from
1986 to 1998 due to poor employees policies .
But this was not the case after 1998 , the retention rate
Of the employees improved drastically from
1999 to 2013

Recommendations

→ Improve the onboarding process.

Maintain a favourable initial impression.

Set clear expectations for their job and the company's future.

Explain what to expect over the first week.

Assist new personnel in integrating into the team and developing relationships with coworkers and colleagues.

Allow new hires to just provide organised feedback on their jobs, corporate processes, and culture.

→Act on insights from exit surveys.

Some employment turnover is unavoidable. Regardless of whether you have a high or low turnover rate, you may learn a lot from employees on their way out the door.

Exit surveys are a great way to obtain direct feedback from your soon-to-be ex-employees about why they're leaving and any suggestions they might have for improving the company. What you learn could surprise you.

Limitations

Some challenges with the formats of the data .

Since the volume of the data is high it is ought to have some null values and errors .