Big Data – Case Study

Subject – Big Data Analytics and Architecture

PROJECT

Data Science Salaries Analysis Project

Data Science Salaries Analysis Project

Project Overview

This project analyses global data science job salaries using the ds_salaries.csv dataset. The dataset provides detailed information about salaries, job titles, experience levels, employment types, company sizes, and remote work ratios for data-related roles across different countries and years.

The goal is to uncover salary trends, remote work impact, and other key insights that reflect the state of the data science job market.

Dataset Description

File Name: ds_salaries.csv
Total Records: 607
Total Columns: 12

Features:

Column Name	Description
work_year	Year in which the salary data was recorded (2020–2022)
experience_level	Level of experience (EN: Entry, MI: Mid, SE: Senior, EX: Executive)
employment_type	Type of employment (FT: Full-time, PT: Part-time, CT: Contract, FL: Freelance)
job_title	Specific job role/title
salary	Raw salary amount in the local currency
salary_currency	Currency type of the salary
salary_in_usd	Salary converted into USD for standard comparison
employee_residence	Country where the employee resides
remote_ratio	Percentage of remote work (0 = On-site, 50 = Hybrid, 100 = Fully remote)
company_location	Location of the employing company
company_size	Size of the company (S = Small, M = Medium, L = Large)

Project Objectives

1. Analyze salary distribution and trends in data-related roles.
2. Compare salaries across experience levels, company sizes, and locations.
3. Study the effect of remote work on salaries.
4. Identify the most common and highest-paying job titles.
5. Provide insights useful for professionals entering or advancing in the data field.

Technologies Used

Tool	Purpose
HiveQL (Apache Hive)	Data querying and aggregation
HDFS / Local Storage	Data storage
Python (Pandas, Matplotlib)	Data analysis and visualization
Excel / CSV	Raw data format
Jupyter Notebook (Optional)	Interactive analysis and documentation

Steps Performed

1. Data Loading:
   Imported ds_salaries.csv into Hive using a CREATE TABLE command with CSV SerDe.
2. Schema Validation:
   Checked column names, data types, and total records.
3. Data Cleaning:
   - Removed unnecessary columns (Unnamed: 0).
   - Verified missing and inconsistent data.
   - Converted salary to numeric type.
4. Exploratory Data Analysis (EDA):
   - Used Hive queries to compute:
     - Average salary by year, experience, and company size.
     - Top-paying job titles and countries.
     - Salary comparison between remote and on-site jobs.
     - Job count distribution by title and employment type.
5. Visualization (Optional in Python):
   - Created bar charts for average salary by experience level and remote ratio.
   - Line charts for salary trends across years.

Key Insights

1. Salary Growth:
   Average salary increased slightly from 2020 to 2022, showing steady growth in the data field.
2. Experience Level Impact:
   Senior (SE) and Executive (EX) professionals earned significantly more than Entry (EN) or Mid (MI) levels.
3. Top Roles:
   - Machine Learning Engineer and Data Scientist were among the highest-paying positions.

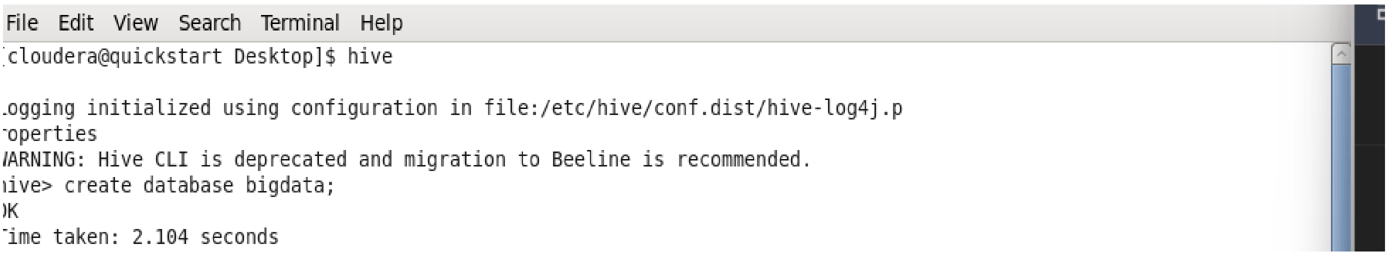
- Data Scientist was also the most common job title (143 occurrences).

4. Remote Work:
   Fully remote (100%) jobs often paid higher than on-site roles, reflecting global flexibility.
5. Company Size:
   Large companies offered slightly higher pay, but medium companies had the most employees.
6. Global Trends:
   - The United States had the majority of employees and top salary ranges.
   - USD was the most common currency used.
7. Employment Type:
   Full-time roles dominated (≈97%), indicating stable, long-term employment trends in the data field.

Conclusion

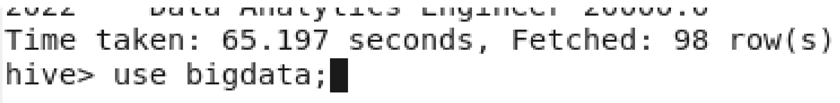
The analysis of the Data Science Salaries dataset reveals a rapidly evolving industry with high-paying opportunities for skilled professionals, particularly in machine learning and data engineering roles. Experience level, company size, and remote flexibility significantly influence salaries.

Create Database:



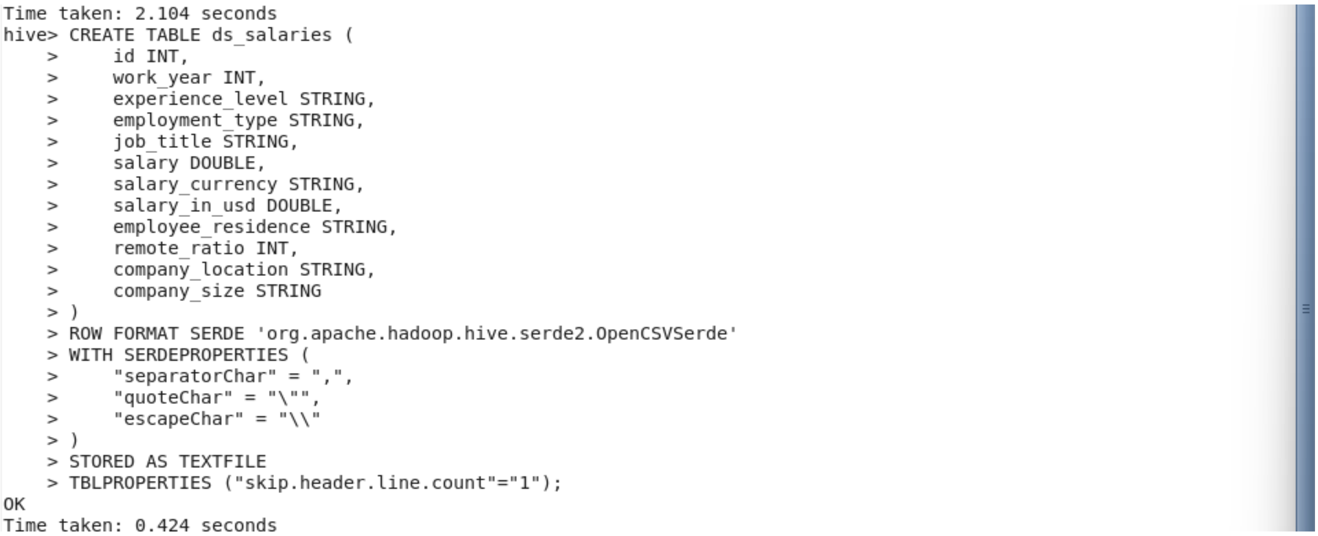
```
[cloudera@quickstart Desktop]$ hive

.ogging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
/ARNING: Hive CLI is deprecated and migration to Beeline is recommended.
ive> create database bigdata;
OK
Time taken: 2.104 seconds
```

Use database:



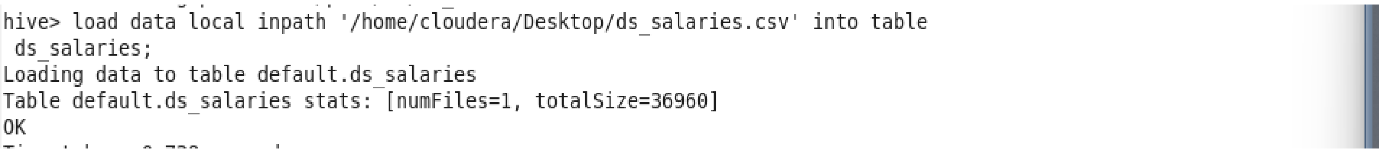
```
2022    Data Analytics Engineer 20000.0
Time taken: 65.197 seconds, Fetched: 98 row(s)
hive> use bigdata;
```

Create Table :



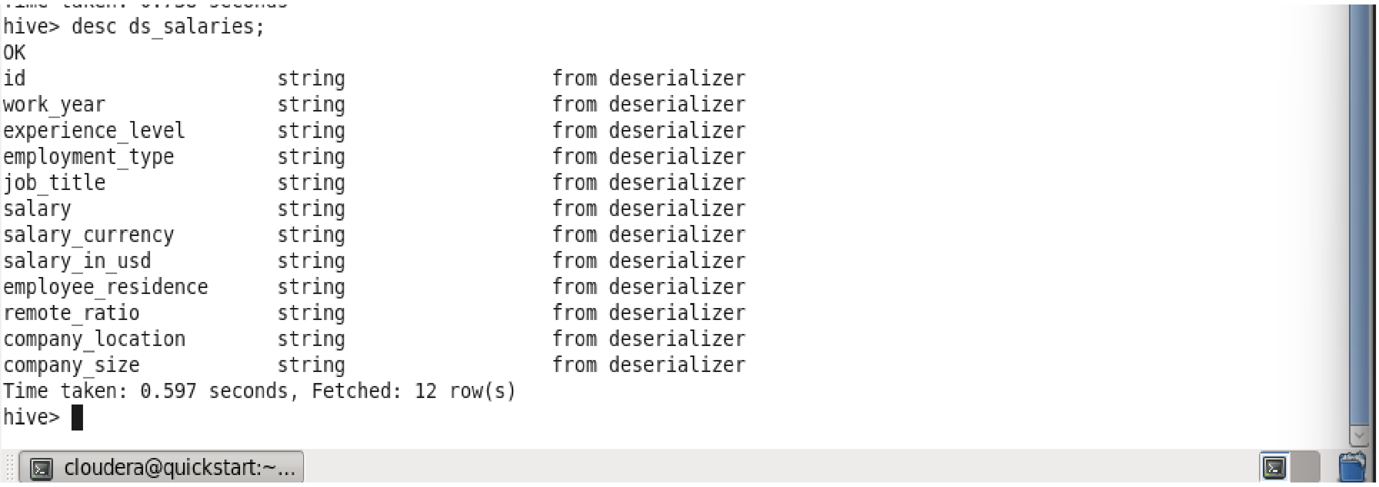
```
Time taken: 2.104 seconds
hive> CREATE TABLE ds_salaries (
    >     id INT,
    >     work_year INT,
    >     experience_level STRING,
    >     employment_type STRING,
    >     job_title STRING,
    >     salary DOUBLE,
    >     salary_currency STRING,
    >     salary_in_usd DOUBLE,
    >     employee_residence STRING,
    >     remote_ratio INT,
    >     company_location STRING,
    >     company_size STRING
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    >     "separatorChar" = ",",
    >     "quoteChar" = "\"",
    >     "escapeChar" = "\\"
    > )
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.424 seconds
```

Load Data:



```
hive> load data local inpath '/home/cloudera/Desktop/ds_salaries.csv' into table
 ds_salaries;
Loading data to table default.ds_salaries
Table default.ds_salaries stats: [numFiles=1, totalSize=36960]
OK
```

Describe Table:

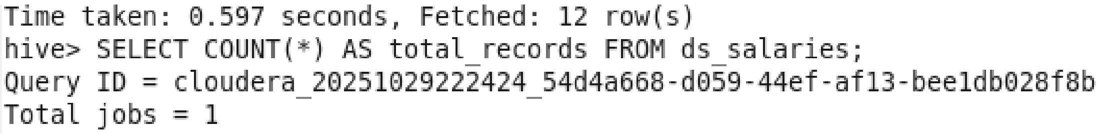


```
hive> desc ds_salaries;
OK
id                      string                  from deserializer
work_year               string                  from deserializer
experience_level        string                  from deserializer
employment_type         string                  from deserializer
job_title               string                  from deserializer
salary                  string                  from deserializer
salary_currency         string                  from deserializer
salary_in_usd           string                  from deserializer
employee_residence      string                  from deserializer
remote_ratio            string                  from deserializer
company_location        string                  from deserializer
company_size            string                  from deserializer
Time taken: 0.597 seconds, Fetched: 12 row(s)
hive> 
```

1. Total number of job records

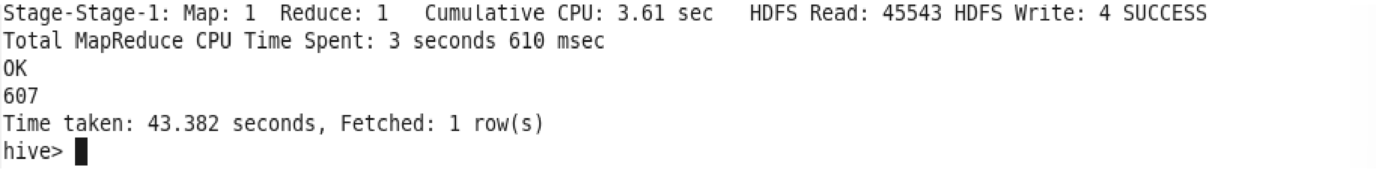
SELECT COUNT(*) AS total_records FROM ds_salaries;

Insight: Total number of salary entries in the dataset.



```
Time taken: 0.597 seconds, Fetched: 12 row(s)
hive> SELECT COUNT(*) AS total_records FROM ds_salaries;
Query ID = cloudera_20251029222424_54d4a668-d059-44ef-af13-bee1db028f8b
Total jobs = 1
```

Output –



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.61 sec   HDFS Read: 45543 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 610 msec
OK
607
Time taken: 43.382 seconds, Fetched: 1 row(s)
hive> 
```

2. Average salary (in USD) by experience level

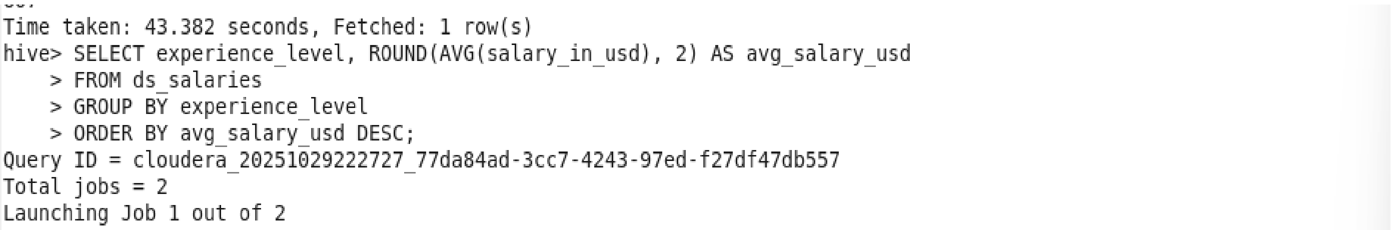
SELECT experience_level, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

GROUP BY experience_level

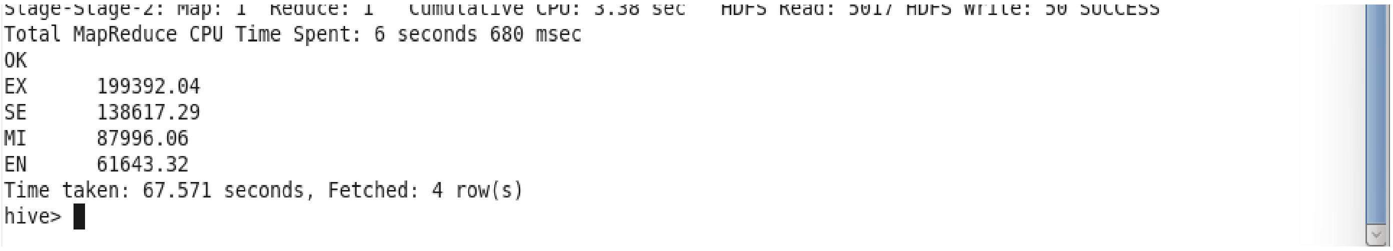
ORDER BY avg_salary_usd DESC;

Insight: Helps identify which experience level (Entry, Mid, Senior, Executive) earns the most.



```
Time taken: 43.382 seconds, Fetched: 1 row(s)
hive> SELECT experience_level, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY experience_level
    > ORDER BY avg_salary_usd DESC;
Query ID = cloudera_20251029222727_77da84ad-3cc7-4243-97ed-f27df47db557
Total jobs = 2
Launching Job 1 out of 2
```

Output –



```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.38 sec   HDFS Read: 5017 HDFS Write: 50 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 680 msec
OK
EX      199392.04
SE      138617.29
MI      87996.06
EN      61643.32
Time taken: 67.571 seconds, Fetched: 4 row(s)
hive> 
```

3. Top 10 highest-paying job titles

SELECT job_title, ROUND (AVG (salary_in_usd), 2) AS avg_salary_usd

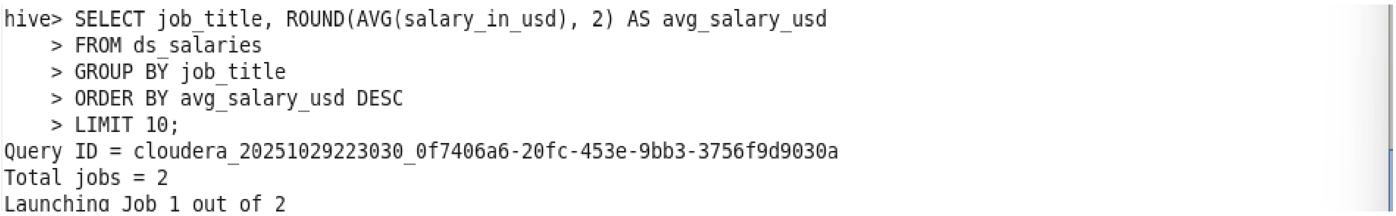
FROM ds_salaries

GROUP BY job_title

ORDER BY avg_salary_usd DESC

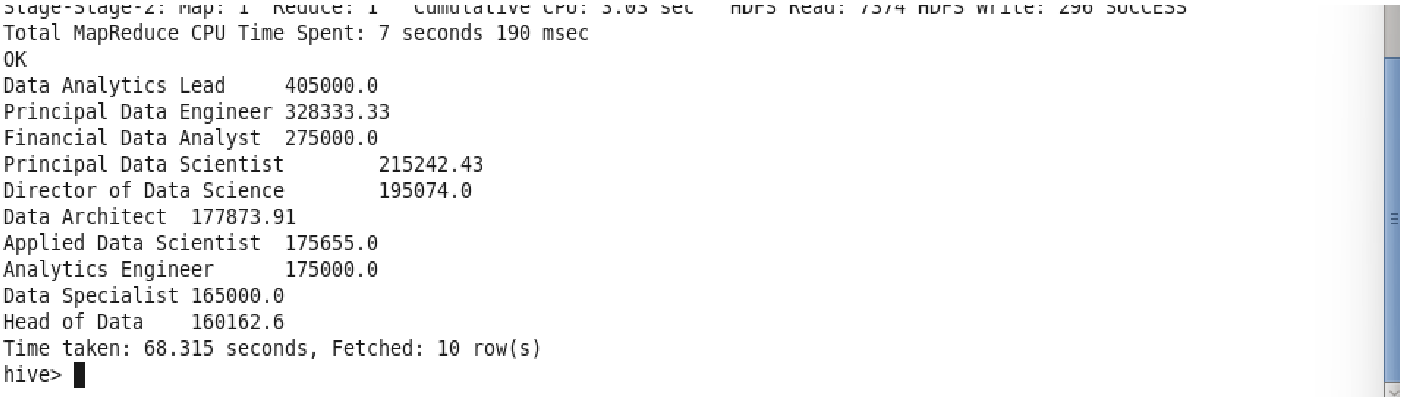
LIMIT 10;

Insight: Reveals which job titles have the highest average pay.



```
hive> SELECT job_title, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY job_title
    > ORDER BY avg_salary_usd DESC
    > LIMIT 10;
Query ID = cloudera_20251029223030_0f7406a6-20fc-453e-9bb3-3756f9d9030a
Total jobs = 2
Launching Job 1 out of 2
```

Output –



```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.03 sec   HDFS Read: 7374 HDFS Write: 296 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 190 msec
OK
Data Analytics Lead     405000.0
Principal Data Engineer 328333.33
Financial Data Analyst  275000.0
Principal Data Scientist        215242.43
Director of Data Science        195074.0
Data Architect  177873.91
Applied Data Scientist  175655.0
Analytics Engineer      175000.0
Data Specialist 165000.0
Head of Data    160162.6
Time taken: 68.315 seconds, Fetched: 10 row(s)
hive> 
```

4. Salary trend by year

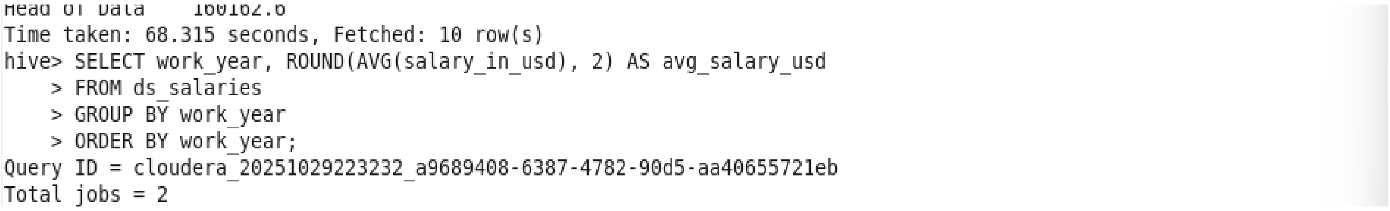
SELECT work_year, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

GROUP BY work_year

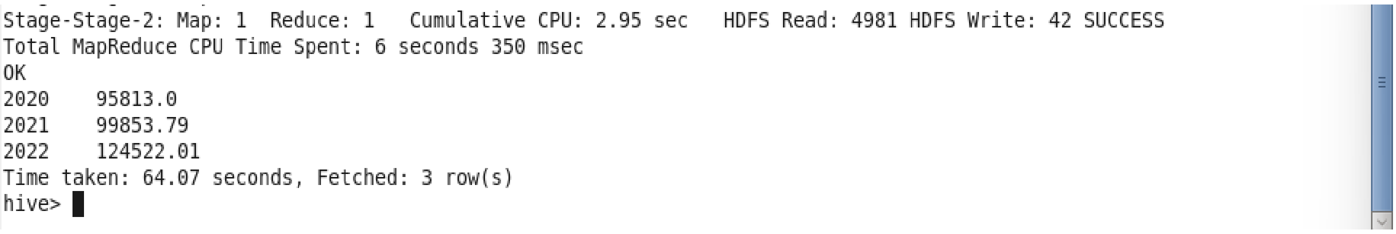
ORDER BY work_year;

Insight: Understands how salaries changed across years (e.g., 2020–2023).



```
Head of Data    160162.6
Time taken: 68.315 seconds, Fetched: 10 row(s)
hive> SELECT work_year, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY work_year
    > ORDER BY work_year;
Query ID = cloudera_20251029223232_a9689408-6387-4782-90d5-aa40655721eb
Total jobs = 2
```

Output –



```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.95 sec   HDFS Read: 4981 HDFS Write: 42 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 350 msec
OK
2020    95813.0
2021    99853.79
2022    124522.01
Time taken: 64.07 seconds, Fetched: 3 row(s)
hive> 
```

5. Average salary by company size

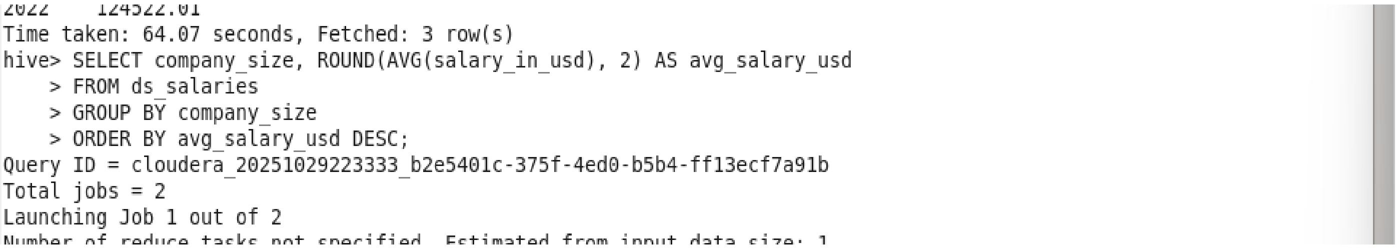
SELECT company_size, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

GROUP BY company_size

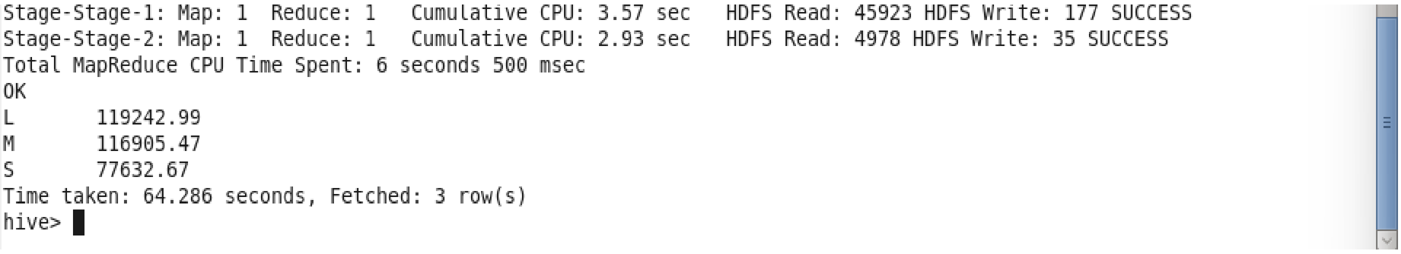
ORDER BY avg_salary_usd DESC;

Insight: Shows if large companies pay more than medium/small ones.



```
2022    124522.01
Time taken: 64.07 seconds, Fetched: 3 row(s)
hive> SELECT company_size, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY company_size
    > ORDER BY avg_salary_usd DESC;
Query ID = cloudera_20251029223333_b2e5401c-375f-4ed0-b5b4-ff13ecf7a91b
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
```

Output –



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.57 sec   HDFS Read: 45923 HDFS Write: 177 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.93 sec   HDFS Read: 4978 HDFS Write: 35 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 500 msec
OK
L       119242.99
M       116905.47
S       77632.67
Time taken: 64.286 seconds, Fetched: 3 row(s)
hive> 
```

6. Remote work impact on salary

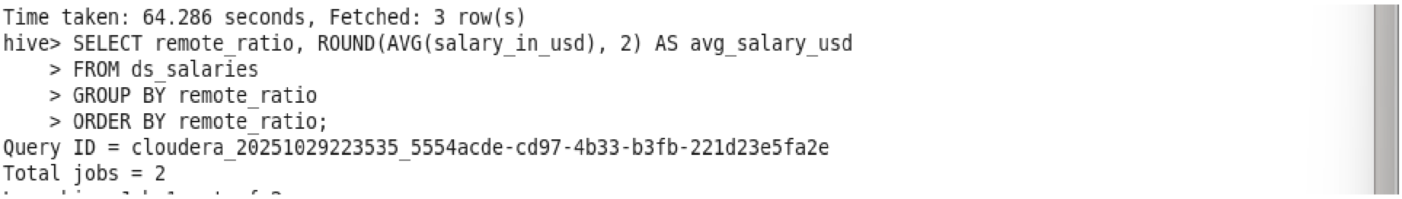
SELECT remote_ratio, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

GROUP BY remote_ratio

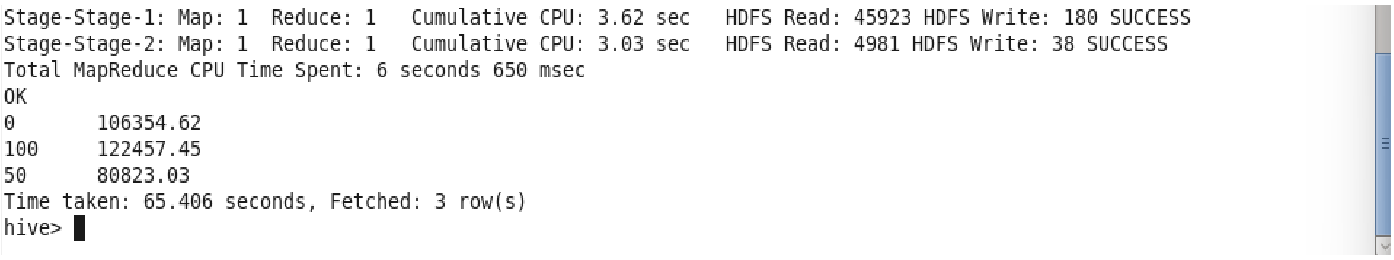
ORDER BY remote_ratio;

Insight: Compare pay between on-site (0%), hybrid (50%), and fully remote (100%) roles.



```
Time taken: 64.286 seconds, Fetched: 3 row(s)
hive> SELECT remote_ratio, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY remote_ratio
    > ORDER BY remote_ratio;
Query ID = cloudera_20251029223535_5554acde-cd97-4b33-b3fb-221d23e5fa2e
Total jobs = 2
```

Output –



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.62 sec   HDFS Read: 45923 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.03 sec   HDFS Read: 4981 HDFS Write: 38 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 650 msec
OK
0       106354.62
100     122457.45
50      80823.03
Time taken: 65.406 seconds, Fetched: 3 row(s)
hive> 
```

7. Top 10 countries with highest average salary

SELECT company_location, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

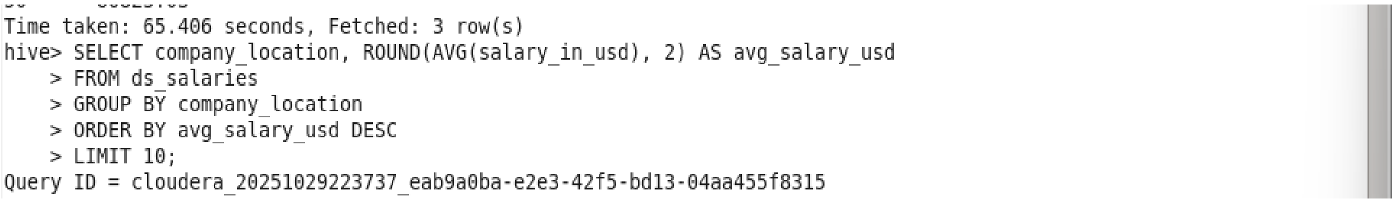
FROM ds_salaries

GROUP BY company_location

ORDER BY avg_salary_usd DESC

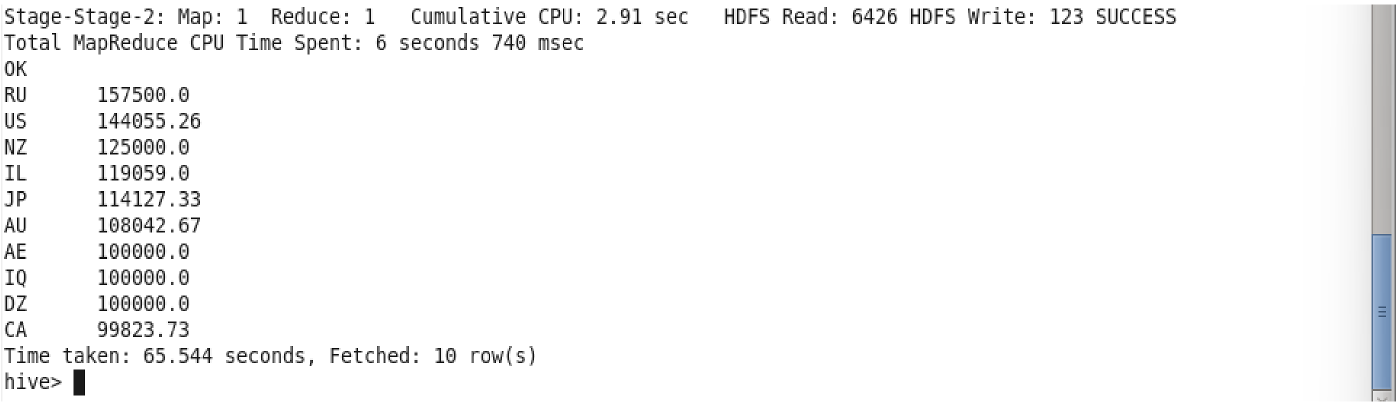
LIMIT 10;

Insight: Find which company locations offer the best pay globally.



```
Time taken: 65.406 seconds, Fetched: 3 row(s)
hive> SELECT company_location, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY company_location
    > ORDER BY avg_salary_usd DESC
    > LIMIT 10;
Query ID = cloudera_20251029223737_eab9a0ba-e2e3-42f5-bd13-04aa455f8315
```

Output –



```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.91 sec   HDFS Read: 6426 HDFS Write: 123 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 740 msec
OK
RU      157500.0
US      144055.26
NZ      125000.0
IL      119059.0
JP      114127.33
AU      108042.67
AE      100000.0
IQ      100000.0
DZ      100000.0
CA      99823.73
Time taken: 65.544 seconds, Fetched: 10 row(s)
hive> 
```

8. Most common job titles

SELECT job_title, COUNT(*) AS job_count

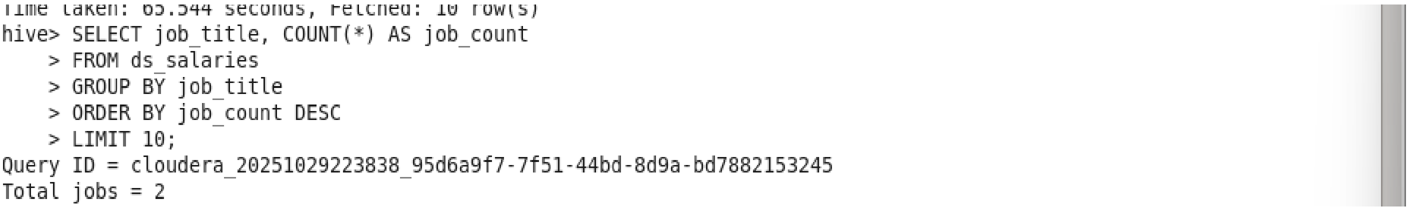
FROM ds_salaries

GROUP BY job_title

ORDER BY job_count DESC

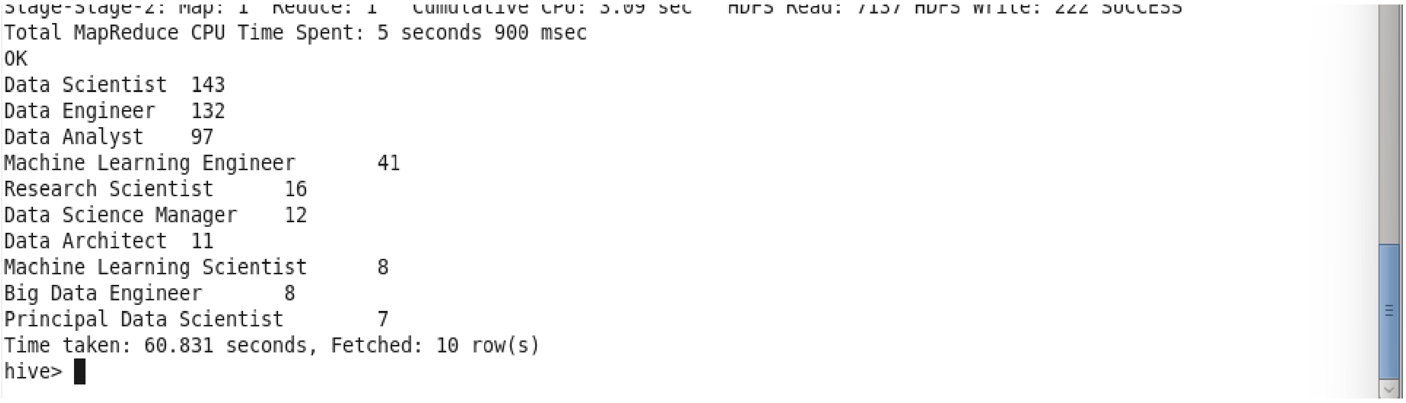
LIMIT 10;

Insight: Shows which roles are most in-demand or frequently listed.



```
Time taken: 65.544 seconds, Fetched: 10 row(s)
hive> SELECT job_title, COUNT(*) AS job_count
    > FROM ds_salaries
    > GROUP BY job_title
    > ORDER BY job_count DESC
    > LIMIT 10;
Query ID = cloudera_20251029223838_95d6a9f7-7f51-44bd-8d9a-bd7882153245
Total jobs = 2
```

Output –



```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.09 sec   HDFS Read: 7137 HDFS Write: 222 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 900 msec
OK
Data Scientist  143
Data Engineer   132
Data Analyst    97
Machine Learning Engineer       41
Research Scientist      16
Data Science Manager    12
Data Architect  11
Machine Learning Scientist      8
Big Data Engineer       8
Principal Data Scientist        7
Time taken: 60.831 seconds, Fetched: 10 row(s)
hive>
```

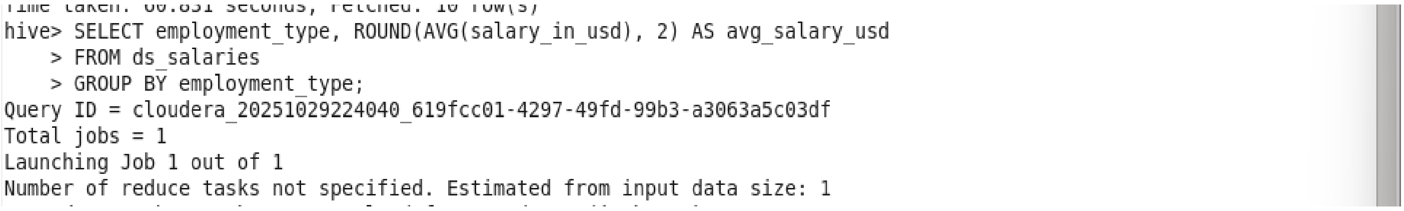
9. Average salary by employment type

SELECT employment_type, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

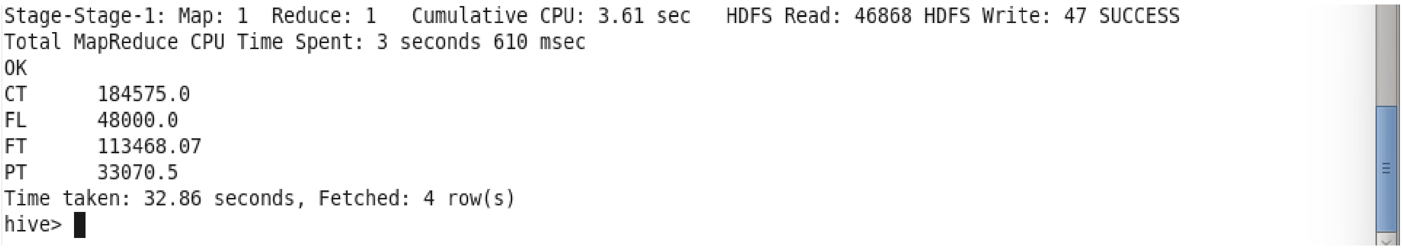
GROUP BY employment_type;

Insight: Compares pay among full-time (FT), part-time (PT), contract (CT), etc.



```
Time taken: 60.831 seconds, Fetched: 10 row(s)
hive> SELECT employment_type, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY employment_type;
Query ID = cloudera_20251029224040_619fcc01-4297-49fd-99b3-a3063a5c03df
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```

Output –



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.61 sec   HDFS Read: 46868 HDFS Write: 47 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 610 msec
OK
CT      184575.0
FL      48000.0
FT      113468.07
PT      33070.5
Time taken: 32.86 seconds, Fetched: 4 row(s)
hive>
```

10. Highest-paid role in each year

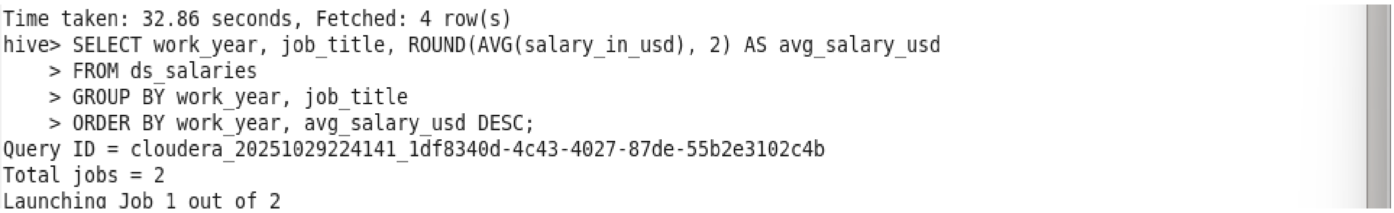
SELECT work_year, job_title, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd

FROM ds_salaries

GROUP BY work_year, job_title

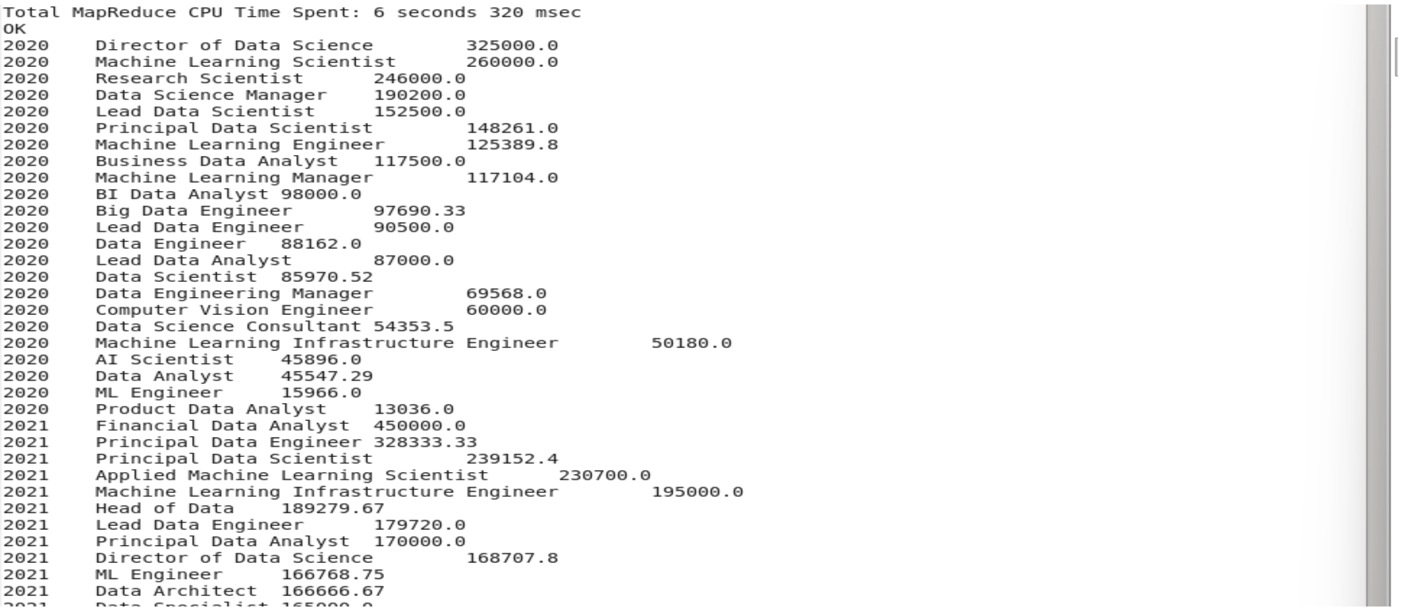
ORDER BY work_year, avg_salary_usd DESC;

Insight: Identifies the top-paying job title for every year.



```
Time taken: 32.86 seconds, Fetched: 4 row(s)
hive> SELECT work_year, job_title, ROUND(AVG(salary_in_usd), 2) AS avg_salary_usd
    > FROM ds_salaries
    > GROUP BY work_year, job_title
    > ORDER BY work_year, avg_salary_usd DESC;
Query ID = cloudera_20251029224141_1df8340d-4c43-4027-87de-55b2e3102c4b
Total jobs = 2
Launching Job 1 out of 2
```

Output –



```
Total MapReduce CPU Time Spent: 6 seconds 320 msec
OK
2020    Director of Data Science        325000.0
2020    Machine Learning Scientist      260000.0
2020    Research Scientist      246000.0
2020    Data Science Manager    190200.0
2020    Lead Data Scientist     152500.0
2020    Principal Data Scientist        148261.0
2020    Machine Learning Engineer       125389.8
2020    Business Data Analyst   117500.0
2020    Machine Learning Manager        117104.0
2020    BI Data Analyst 98000.0
2020    Big Data Engineer       97690.33
2020    Lead Data Engineer      90500.0
2020    Data Engineer   88162.0
2020    Lead Data Analyst       87000.0
2020    Data Scientist  85970.52
2020    Data Engineering Manager        69568.0
2020    Computer Vision Engineer        60000.0
2020    Data Science Consultant 54353.5
2020    Machine Learning Infrastructure Engineer        50180.0
2020    AI Scientist    45896.0
2020    Data Analyst    45547.29
2020    ML Engineer     15966.0
2020    Product Data Analyst    13036.0
2021    Financial Data Analyst  450000.0
2021    Principal Data Engineer 328333.33
2021    Principal Data Scientist        239152.4
2021    Applied Machine Learning Scientist      230700.0
2021    Machine Learning Infrastructure Engineer        195000.0
2021    Head of Data    189279.67
2021    Lead Data Engineer      179720.0
2021    Principal Data Analyst  170000.0
2021    Director of Data Science        168707.8
2021    ML Engineer     166768.75
2021    Data Architect  166666.67
2021    Data Specialist 165000.0
```