Import Libraries

```
!apt install -qq -y -q -w -t -f
!apt-get install -qq -y -q -w -t -f
!apt-get update
```

```
[sandeep@ip-172-31-16-234 ~]$ ls -ltrh /home/ec2-user/*.csv
-rw-rw-r-- 1 ec2-user ec2-user 1.1G Jun 13 18:15 /home/ec2-user/cleaned_data.csv
[sandeep@ip-172-31-16-234 ~]$ head -1 /home/ec2-user/cleaned_data.csv
"job_title","job_description","company_name","location","salary_estimate","company_rating","company_size","company_type","industry","sector","revenue","competitors","job_posted_date","job_posted_year","job_posted_month","job_posted_day","job_posted_weekday","job_posted_hour","job_posted_minute","job_posted_second","job_posted_timestamp","job_posted_unix_timestamp","salary_lower","salary_upper","salary_mean"
```

Writing records to JSON file

```
[sandeep@ip-172-31-16-234 ~]$ /home/ec2-user/spark-3.5.1-bin-hadoop3/bin/spark-shell --executor-memory 4G --driver-memory 4G
WARNING: Use "yarn jar" to launch YARN applications.
Setting default log level to "WARN".
```

Merging skills files

```
scala> cleaned_skills.coalesce(1).write.mode("overwrite").parquet("/home/ec2-user/cleaned_skills.parquet")
[Stage 0:>(0 + 1) / 1]
[Stage 1:>(0 + 1) / 1]
[Stage 2:>(0 + 1) / 1]
[Stage 3:>(0 + 1) / 1]
[Stage 4:>(0 + 1) / 1]
[Stage 5:>(0 + 1) / 1]
[Stage
```

Category	Value	Unit/Key
Rows	2,253	Data Analyst
Rows	2,253	Data Scientist
Rows	315	Data Analyst
Rows	315	Data Scientist
Rows	1,938	Data Analyst
Rows	1,938	Data Scientist

```
[ 2253, 2253, 315, 315, 1938, 1938 ]
```

Title	id	description	date_posted	location_id	link	seniority_level	employment
Data Analyst, Center for Biological Diversity	4.0E+09	2022-01-06	2022-01-06	2151	https://www.linkedin.com/jobs/view/3526...?refId=...&trk=jobs_jobs-search-job-card-title	Entry level	Full-time