Kevin Kao
Harshavardhan Suresh
Divya Vishnumurthy

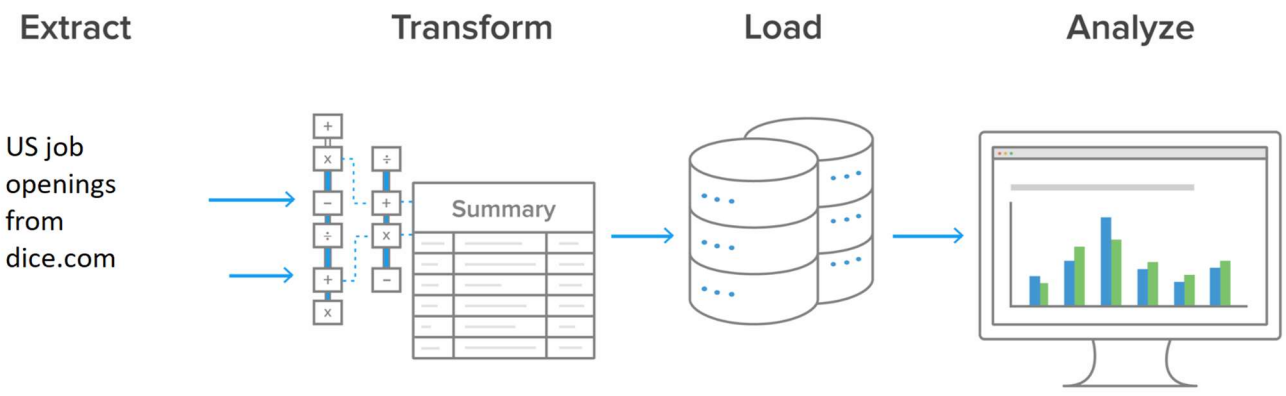
ETL Project

US Technology jobs, 2019

ETL is an important part of today's business intelligence (BI) processes and systems. It is the IT process from which data from disparate sources can be put in one place to programmatically analyze and discover business insights.

In our project we were interested in understanding the job market for data analysts. We pulled data from dice.com where all the technology job openings are posted. We then wanted to check if a certain city, state had more requirements than others. Also, what are the top three skills that companies are looking for. Our transformed and loaded data can also answer questions like: if there is a relation between job title and travel. If a job title allows to telecommute and what % of time, list of companies that look for certain skills and where they are located.

We used both SQL and NOSQL Databases to load our data based on the data type we were transforming.



01

Data Extraction: Kaggle dataset- US Technology job openings from dice.com

02

Data Transformation: Python pandas and dataframe, dictionaries and string manipulations.

03

Data Loading: SQLAlchemy, pymongo, SQL-Postgres and NOSQL-MongoDB.

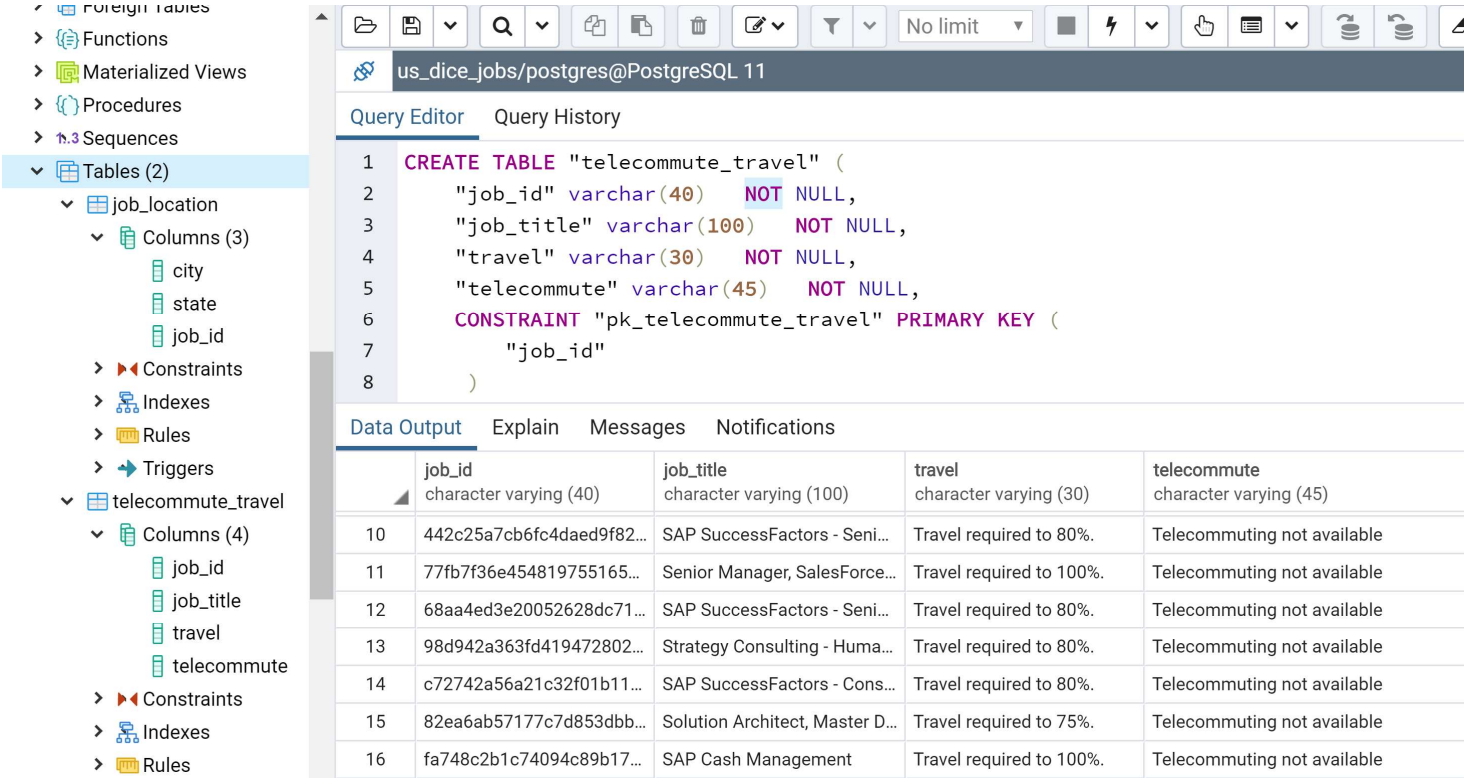
About the Dataset

This is a pre-crawled dataset, taken as subset of a bigger dataset (more than 4.6 million job listings) that was created by extracting data from Dice.com, a prominent US-based technology job board. The data set contains 22K records.

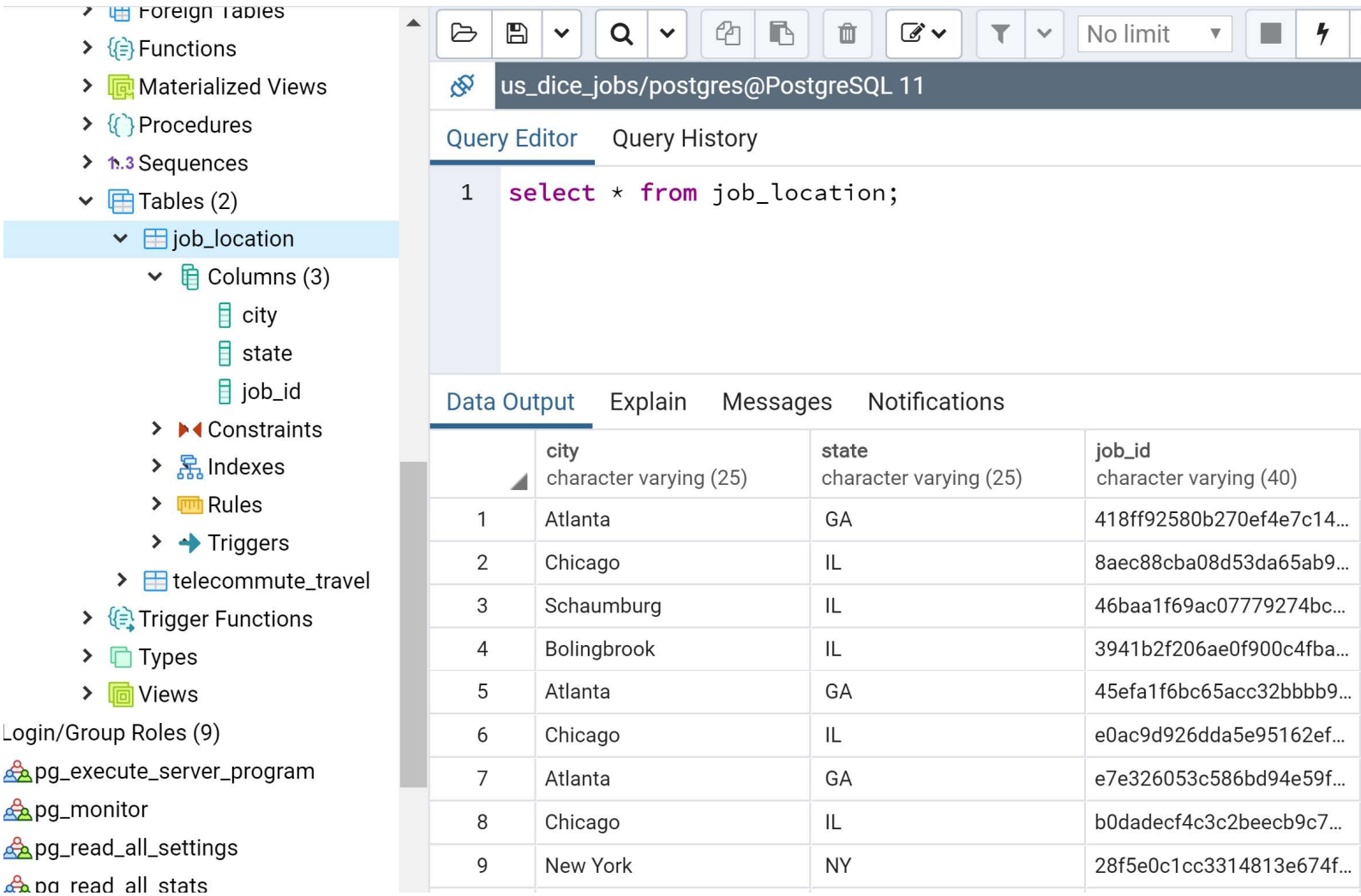
Columns in the dataset	Fields of interest
advertiserurl company employmenttype_jobstatus jobdescription jobid joblocation_address jobtitle postdate shift site_name skills uniq_id	company jobdescription joblocation_address jobtitle shift skills uniq_id

Challenges

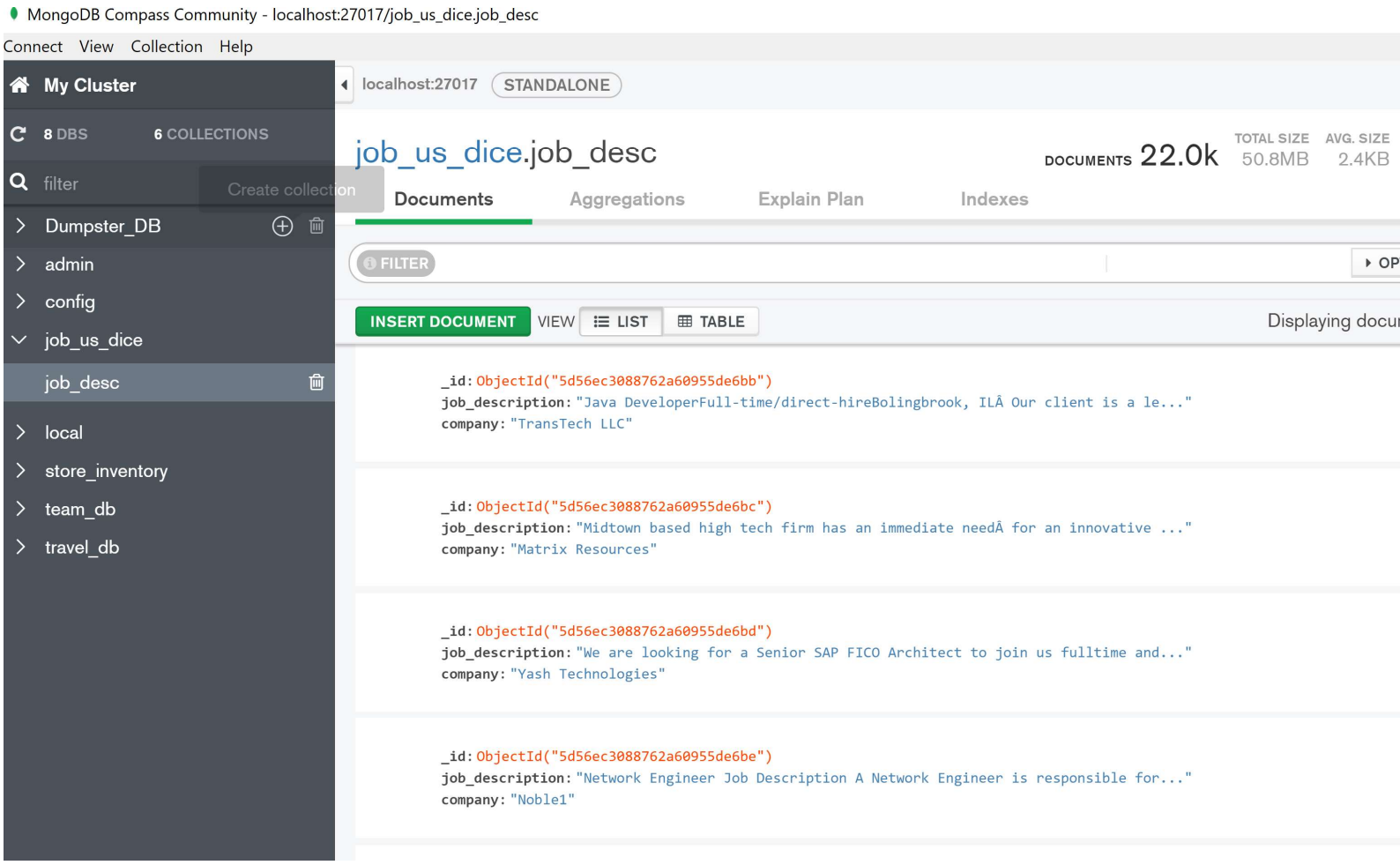
1. Find a dataset which fit the requirements.
2. The irrelevant data had to be filtered out as needed and the comma separated skills had to be split.
3. Jobaddress_location – Contained multiple elements separated by a comma and in some records, they had entire address. We had to clean and transform the data into two columns containing city and state only.
4. Shift column- Contained multiple elements separated by a "|" and some rows were empty. We had to clean those and transform them into two dataframe columns – Travel and Telecommute.
5. Job Description column- Contained a paragraph of the job requirement and this field is more like a document. We tried to fit it onto SQLDB, but it fit better in NOSQL DB.
6. The skills data was gnarly and did not have the values in a uniform format. While the column contained skills as comma separated values, there were also values that were either referencing columns or contained the job description rather than the skills required.
7. Once the data was cleaned up, a judgement call had to be made to confirm on the type of database (SQL vs NOSQL). Given that the number of skills varied, we chose mongo DB, NOSQL as it was easier to reference the skills and pull out the relevant job details.

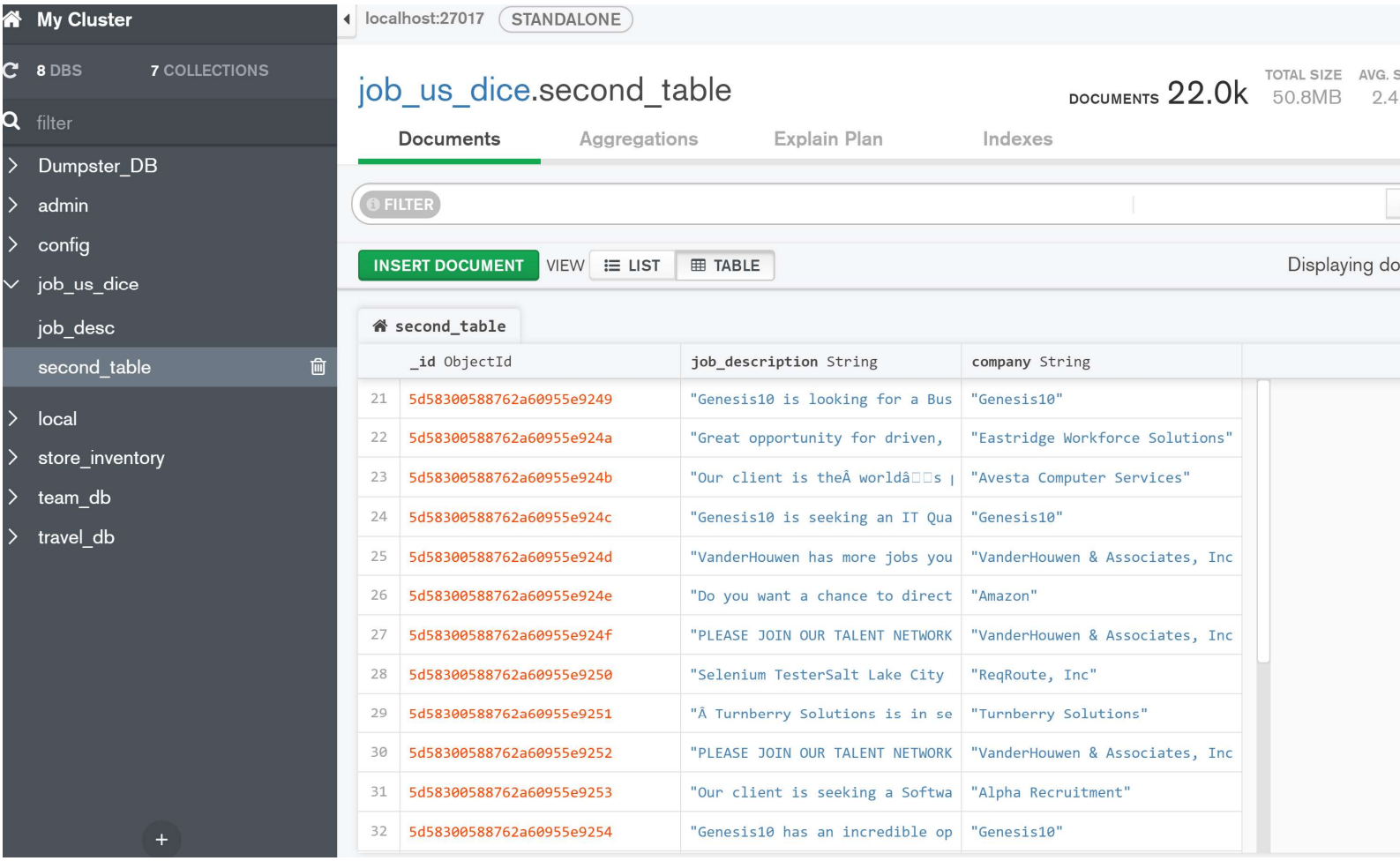


	job_id character varying (40)	job_title character varying (100)	travel character varying (30)	telecommute character varying (45)
10	442c25a7cb6fc4daed9f82...	SAP SuccessFactors - Seni...	Travel required to 80%.	Telecommuting not available
11	77fb7f36e454819755165...	Senior Manager, SalesForce...	Travel required to 100%.	Telecommuting not available
12	68aa4ed3e20052628dc71...	SAP SuccessFactors - Seni...	Travel required to 80%.	Telecommuting not available
13	98d942a363fd419472802...	Strategy Consulting - Huma...	Travel required to 80%.	Telecommuting not available
14	c72742a56a21c32f01b11...	SAP SuccessFactors - Cons...	Travel required to 80%.	Telecommuting not available
15	82ea6ab57177c7d853dbb...	Solution Architect, Master D...	Travel required to 75%.	Telecommuting not available
16	fa748c2b1c74094c89b17...	SAP Cash Management	Travel required to 100%.	Telecommuting not available



	city character varying (25)	state character varying (25)	job_id character varying (40)
1	Atlanta	GA	418ff92580b270ef4e7c14...
2	Chicago	IL	8aec88cba08d53da65ab9...
3	Schaumburg	IL	46baa1f69ac07779274bc...
4	Bolingbrook	IL	3941b2f206ae0f900c4fba...
5	Atlanta	GA	45efa1f6bc65acc32bbbb9...
6	Chicago	IL	e0ac9d926dda5e95162ef...
7	Atlanta	GA	e7e326053c586bd94e59f...
8	Chicago	IL	b0dadecf4c3c2beecb9c7...
9	New York	NY	28f5e0c1cc3314813e674f...


MongoDB Compass Community - localhost:27017/job_us_dice.job_desc
Connect View Collection Help
My Cluster
8 DBS 6 COLLECTIONS
filter
Create collection
Dumpster_DB
admin
config
job_us_dice
job_desc
local
store_inventory
team_db
travel_db
localhost:27017 STANDALONE
job_us_dice.job_desc
DOCUMENTS 22.0k
TOTAL SIZE 50.8MB
AVG. SIZE 2.4KB
Documents
Aggregations
Explain Plan
Indexes
FILTER
INSERT DOCUMENT
VIEW LIST TABLE
Displaying docu
_id: ObjectId("5d56ec3088762a60955de6bb")
job_description: "Java DeveloperFull-time/direct-hireBolingbrook, ILÂ Our client is a le..."
company: "TransTech LLC"
_id: ObjectId("5d56ec3088762a60955de6bc")
job_description: "Midtown based high tech firm has an immediate needÂ for an innovative ..."
company: "Matrix Resources"
_id: ObjectId("5d56ec3088762a60955de6bd")
job_description: "We are looking for a Senior SAP FICO Architect to join us fulltime and..."
company: "Yash Technologies"
_id: ObjectId("5d56ec3088762a60955de6be")
job_description: "Network Engineer Job Description A Network Engineer is responsible for..."
company: "Noble1"



us_jobs/postgres@PostgreSQL 11

Query Editor | Query History | Scratch Pad

```
1  select * from dice_jobs
```

Data Output | Explain | Messages | Notifications

	ID integer	date_added date	job_title character varying	job_type character varying	location character varying	organization character varying	sector character varying
1	0	2016-11-11	EDI Analyst	Full Time, Full-time, Employee	Stamford, CT	CyberCoders	EDI, TrustedLink, AS2, VAN - EDI, Trust
2	1	2016-11-11	Informatica ETL Developer	Full Time, Full Time	St Petersburg, FL	TrustMinds	ETL Informatica B2B Data Exchange N
3	2	2016-11-11	Angular developer	Full Time, Contract Corp-To-Corp, Con...	Sunnyvale, CA	K Anand Corporation	Angular
4	3	2016-11-12	Microsoft Dynamics AX, Project Manager	Full Time	Toronto, Canada, ON	Nigel Frank International	Microsoft Dynamics AX, Project Mana
5	4	2016-11-11	Software Developer	Full Time, Full-time, Employee	Stamford, CT	CyberCoders	C#, ASP.NET, SQL, JavaScript, MVC - C
6	5	2016-11-11	Machine Learning Engineer	Full Time, Full-time, Employee	Toronto, ON	CyberCoders	Machine Learning, Java/ Scala, SPARK
7	6	2016-11-11	Core Java Developer	Full Time, C2H W2, FTE - C2H	Stamford, CT	Hatstand US	Java, Linux/Unix, SDLC; Multi-Threade
8	7	2016-11-11	Linux System Administrator	Full Time	Stamford, CT	Landover Assocates	Linux System Administrator
9	8	2016-11-11	Commission Analyst	Full Time, Contract Independent, Contr...	Sunnyvale, CA	Varite	Commission Analyst, Sales commissio
10	9	2016-11-12	Drupal Developer	Contract Corp-To-Corp, Contract Indep...	Sunnyvale, CA	Tranzeal	strong experience in Drupal applicatio
11	10	2016-11-12	SQL/SSIS Developer (AWS Migration Project)	Contract Corp-To-Corp, Contract Indep...	Santa Monica, CA	Tentek	SSIS, AWS, Scripting Tool
12	11	2016-11-11	PMO (for Purchase to Pay Support) OR Pur...	Full Time, Contract Corp-To-Corp, Con...	Sunnyvale, CA	Brainhunter Companies LLC	Purchase to Pay Support, PMO, Retail
13	12	2016-11-11	Senior .NET Developer	Full Time, Full-time, Employee	Stamford, CT	CyberCoders	C#, ASP.NET, JavaScript, Content Mana
14	13	2016-11-12	Field Service Technician II US	Contract Independent, Contract W2, C...	Seattle, WA	Matrix Resources	Hardware, Laptop, Networking, Periphe
15	14	2016-11-12	Software Development Engineer	Full Time	Seattle, WA	Amazon	C++, Development, Ecommerce, Java,
16	15	2016-11-12	Business Operations /Financial Analyst	Contract Corp-To-Corp, Contract Indep...	Sunnyvale, CA	Opal Soft	BI, ad hoc reports, data analysis
17	16	2016-11-11	Personalization Specialist	Contract W2	55 Bloor St W, Toronto, Ontario, Canada, ON	TEKsystems	Personalization Specialist
18	17	2016-11-11	SAP Ariba Consultant (Supply Chain Capab...	Full Time, Permanent	Tallahassee, FL	Deloitte	Ariba, Consulting, CRM, Development,
19	18	2016-11-12	MONGOdb and JAVA	Contract Corp-To-Corp, Contract Indep...	Sunnyvale, CA	ZealTech	Mongodb java
20	19	2016-11-12	Software Engineer	Full Time	Tampa, FL	Ciber	QUALIFICATIONS: * Experience workin
21	20	2016-11-12	System Administrator	C2H Corp-To-Corp, C2H Independent	Tampa, FL	Robert Half Technology	Analysis, Backup and Recovery, CCNA