Database Systems for Analytics

Project Report

Shahraiz Qureshi 009551906 Jennifer Kwok 010395853 Gangotri Biswal 015320630
Palak Shah 015272803 Seung min Yoo 015280772 Chitra Priyaa 014504243

Abstract - The project aims at professionals who are looking for a job as a Data Analyst. Amidst the pandemic, many people lost their jobs, with this dataset it is possible to hone the job search so that more people in need can find employment. This dataset was created by *picklesueat* and contains more than 2000 job listings for data analyst positions with features such as the Salary Estimate, Location, Company Rating, Job Description.

Index Terms - DataAnalyst, Cassandra, No-SQL, DataStax Enterprise (DSE), DataStax Studio, ETL, CQL / SQL, JSON, Apache Solr, Dsbulkloadxr, Knowi

1 INTRODUCTION

The analytics field is on the rise worldwide. In this project, we will take a closer look at the US job market and figure out some facts about DA jobs. This analysis will help us to understand different features like which keywords are used in job titles for data analyst roles and which industry posts maximum data analyst jobs, etc. Dataset from Kaggle was used and the ETL process along with other software was used to analyze and visualize data. Cassandra is a NoSQL Database Management distributed database system which means it's available on multiple computers at once. The user can talk to one computer, get data from that computer, then tell a different computer to get different data or possibly the same data from that different computer. Cassandra is widely used in tech firms, Facebook being the pioneer which was the main company that developed Cassandra DB. It has fast writes meaning the user can write a lot of data to the database quickly. It is used in the big data world for analytics, internet of things IOT where there are a lot of devices writing a lot of data very quickly. It is very useful for time series data; it can be very quick and very efficient in those situations. Cassandra takes a column-oriented approach to storing and querying data compared to traditional databases such as MySQL, Postgres which are row-oriented approaches.

2 METHODOLOGY

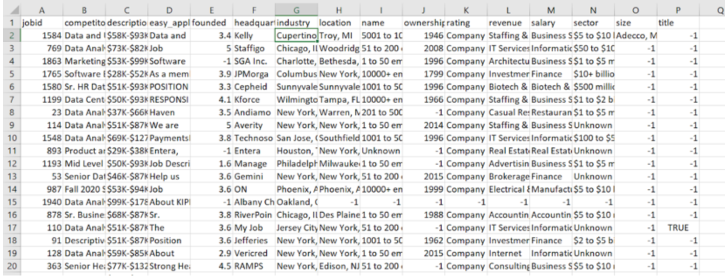
The unique feature of Cassandra is it stores the databases in columns, while a traditional database will store them in a row. Column-oriented databases store each column on a separate file on a disk giving the advantage of just accessing the columns that the user is looking for in a query. Cassandra is a distributed database which means it's available on multiple computers at once. The user can talk to one computer, get data from that computer, then tell a different computer to get different data or possibly the same data from that different computer. Cassandra is widely used in tech firms Facebook being the pioneer which was the main company that developed Cassandra DB. It has fast writes meaning the user can write a lot of data to the database quickly. It is used in the big data world for analytics, internet of things IOT where there are a lot of devices writing a lot of data very quickly. It is very useful for time series data; it can be very quick and very efficient in those situations. Cassandra takes a column oriented approach to storing and querying data compared to traditional databases such as MySQL ,Postgres which are row oriented approaches. The unique feature of Cassandra is it stores the databases in columns, while a traditional database will store it them in a row. Column oriented databases store each column on a separate file on a disk giving the advantage of just accessing the columns that the user is looking for in a query. For example, if the user is interested in data contained.

in 2 columns then it only must access those 2 specific columns. Compared to a traditional relational database if the user wants to access 2 specific attributes stored in 2 specific columns it will go through every row including the data the user is not concerned with. In a column-oriented database like Cassandra, it's like a two-dimensional key value pair. Each row will have it's own ID and each column within the row has a key and a pair value. Being able to access just the data or columns the user is interested in speeds up the query duration. It also helps when data needs to be compacted. In a relational database, every row and column should be filled with either null or void statements meaning something is being stored on disk and filling it up with basically nothing. In Cassandra column- oriented databases not every space has to be filled resulting in fewer things stored on memory. It is fault tolerant meaning in multiple nodes if a single node or computer fails the user can still access the data they won't ever lose data if part of the network is down, the user can still operate and run queries. It has great performance, it is superfast, decentralized meaning there is no single point of failure, meaning if multiple nodes are set up and one goes down the user can still operate there is no reliant on a master node controlling everything. Every node is independent of each other. It is scalable for example apples has over 75k nodes storing over 10PB of data and it can keep adding more nodes to increase it's capacity. Cassandra is durable just like fault tolerance the data is always persisted and elastic, meaning read and write throughput both increases linearly. When new machines are added to the cluster the user is getting the benefit of the whole new machine. In Cassandra, the user will always see the elastic benefit of every machine. Cassandra is built on the basis of CAP theorem, which is the backbone theorem of most distributed system applications and distributed databases. CAP theorem is also known as Brewer's theorem; the theorem relates to distributed databases or a system can have. CAP stands for three properties a distributed system can have available, consistent, and partition tolerant. The theorem states that at any one time a system can have only two of the three things stated above, it cannot have three (availability, consistency, partition tolerance). Cassandra cannot be consistent, available, and partition tolerant at the same time. A Consistent database is when every node always returns the same most recently written data, meaning if the user decides to change some attributes of the saved data all the systems will return the latest changed attribute vs a non-consistent DB it might return the previous entry instead of the latest modified entry. Then there's availability meaning every non-failing node returns a response to any read or write request in a reasonable period. The statement means if the user wants to give or get it should always be able to execute that. The DB shout always is available to give the data the user wants or allow the user to write the data to the database that they want. Lastly, it is partition tolerant meaning the system will continue to function even during a network partition or failure. A network partition occurs when some of the nodes cannot communicate with each other. In most distributed systems network partitions, will always occur. In the end, the choice comes to availability for consistency is the decision the user should make. On the grand scheme of database comparison, Cassandra leans more towards the availability attribute. This does not mean Cassandra has bad consistency. The consistency is still good level depending on cluster set up. This can be seen reflected in brewer's revision of the cap theorem in 2012, which states the system can have a lot of availability and a lot of consistency it doesn't always have to be completely warned and none of the other it can be a bit of both, and Cassandra let the user achieve that through several of it's feature. Particularly when the user is reading and writing data they can choose the level of consistency and set the replication factor of the cluster, which is how much node the data is replicated to[7].DataStax Enterprise (DSE) was the data management addon used to do the ETL process. DataStax Bulk loader was used to load data to the cluster. Then CQL was used to run queries. All this was virtualized using Docker to run containers of packages required for Cassandra DB to operate.

2.1 Dataset

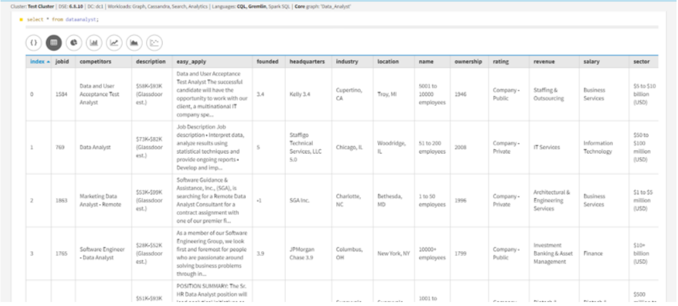
This dataset was created by picklesueat [6] and it contains more than 2000 job listings for data analyst positions, with features such as Salary Estimate, Location, Company Rating, Job Description and more.

It can be used to answer important questions like 'What are the best jobs by salary and company rating', 'Top 5 state of highest average salary', 'What are the Top 20 cities with their minimum and maximum salaries'.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	jobid	competito	descriptio	easy_appl	founded	headquar	industry	location	name	ownershi	rating	revenue	salary	sector	size	title	
2	1584	Data and I	$58K-$93K	Data and	3.4	Kelly	Cupertino	Troy, MI	5001 to 10	1946	Company	Staffing &	Business S	$5 to $10 I	Adecco, M	-1	
3	769	Data Anal	$73K-$82K	Job	5	Staffigo	Chicago, IL	Woodridg	51 to 200	2008	Company	IT Services	Informatic	$50 to $10	-1	-1	
4	1863	Marketing	$53K-$99K	Software	-1	SGA Inc.	Charlotte,	Bethesda,	1 to 50 em	1996	Company	Architectu	Business S	$1 to $5 m	-1	-1	
5	1765	Software E	$28K-$52K	As a mem	3.9	JPMorga	Columbus	New York,	10000+ en	1799	Company	Investmer	Finance	$10+ billio	-1	-1	
6	1580	Sr. HR Dat	$51K-$93K	POSITION	3.3	Cepheid	Sunnyvale	Sunnyvale	1001 to 50	1996	Company	Biotech &	Biotech &	$500 millio	-1	-1	
7	1199	Data Cent	$50K-$93K	RESPONSI	4.1	Kforce	Wilmingto	Tampa, FL	10000+ en	1966	Company	Staffing &	Business S	$1 to $2 bi	-1	-1	
8	23	Data Anal	$37K-$66K	Haven	3.5	Andiamo	New York,	Warren, N	201 to 500	-1	Company	Casual Re	Restauran	$1 to $5 m	-1	-1	
9	114	Data Anal	$51K-$87K	We are	5	Averity	New York,	New York,	1 to 50 em	2014	Company	Staffing &	Business S	Unknown	-1	-1	
10	1548	Data Anal	$69K-$127	Payments	3.8	Technoso	San Jose, (	Southfield	1001 to 50	1996	Company	IT Services	Informatic	$100 to $5	-1	-1	
11	893	Product ar	$29K-$38K	Entera,	-1	Entera	Houston, T	New York,	Unknown	-1	Company	Real Estat	Real Estat	Unknown	-1	-1	
12	1193	Mid Level	$50K-$93K	Job Descri	1.6	Manage	Philadelph	Milwauke	1 to 50 em	-1	Company	Advertisin	Business S	$1 to $5 m	-1	-1	
13	53	Senior Dat	$46K-$87K	Help us	3.6	Gemini	New York,	New York,	51 to 200	2015	Company	Brokerage	Finance	Unknown	-1	-1	
14	987	Fall 2020 S	$53K-$94K	Job	3.6	ON	Phoenix, A	Phoenix, A	10000+ en	1999	Company	Electrical &	Manufactu	$5 to $10 I	-1	-1	
15	1940	Data Anal	$99K-$178	About KIP	-1	Albany Ch	Oakland, (	-1	-1	-1	-1	-1	-1	-1	-1	-1	
16	878	Sr. Busine	$68K-$87K	Sr.	3.8	RiverPoin	Chicago, IL	Des Plaine	1 to 50 em	1988	Company	Accountin	Accountin	$5 to $10	-1	-1	
17	110	Data Anal	$51K-$87K	The	3.6	My Job	Jersey City	New York,	51 to 200	-1	Company	IT Services	Informatic	Unknown	-1	TRUE	
18	91	Descriptiv	$51K-$87K	Position	3.6	Jefferies	New York,	New York,	1001 to 50	1962	Company	Investmer	Finance	$2 to $5 bi	-1	-1	
19	128	Data Anal	$59K-$85K	About	2.9	Vericred	New York,	New York,	1 to 50 em	2015	Company	Internet	Informatic	Unknown	-1	-1	
20	363	Senior He	$77K-$132	Strong He	4.5	RAMPS	New York,	Edison, NJ	51 to 200	-1	Company	Consulting	Business S	$5 to $10	-1	-1	

Figure 1. Shows the raw data that refined and cleaned



Cluster: **Test Cluster** | DSE: **6.8.10** | DC: dc1 | Workloads: Graph, Cassandra, Search, Analytics | Languages: **CQL, Gremlin**, Spark SQL | **Core** graph: 'Data_Analyst'

```
select * from dataanalyst;
```

index ▲	jobid	competitors	description	easy_apply	founded	headquarters	industry	location	name	ownership	rating	revenue	salary	sector
0	1584	Data and User Acceptance Test Analyst	$58K-$93K (Glassdoor est.)	Data and User Acceptance Test Analyst The successful candidate will have the opportunity to work with our client, a multinational IT company spe...	3.4	Kelly 3.4	Cupertino, CA	Troy, MI	5001 to 10000 employees	1946	Company - Public	Staffing & Outsourcing	Business Services	$5 to $10 billion (USD)
1	769	Data Analyst	$73K-$82K (Glassdoor est.)	Job Description Job description • Interpret data, analyze results using statistical techniques and provide ongoing reports • Develop and imp...	5	Staffigo Technical Services, LLC 5.0	Chicago, IL	Woodridge, IL	51 to 200 employees	2008	Company - Private	IT Services	Information Technology	$50 to $100 million (USD)
2	1863	Marketing Data Analyst - Remote	$53K-$99K (Glassdoor est.)	Software Guidance & Assistance, Inc., (SGA), is searching for a Remote Data Analyst Consultant for a contract assignment with one of our premier fi...	-1	SGA Inc.	Charlotte, NC	Bethesda, MD	1 to 50 employees	1996	Company - Private	Architectural & Engineering Services	Business Services	$1 to $5 million (USD)
3	1765	Software Engineer - Data Analyst	$28K-$52K (Glassdoor est.)	As a member of our Software Engineering Group, we look first and foremost for people who are passionate around solving business problems through in...	3.9	JPMorgan Chase 3.9	Columbus, OH	New York, NY	10000+ employees	1799	Company - Public	Investment Banking & Asset Management	Finance	$10+ billion (USD)
			$51K-$93K	POSITION SUMMARY: The Sr. HR Data Analyst position will					1001 to					$500

Figure 2. This shows the data when it was bulk loaded in Cassandra DB using DSbulkload

2.2 Framework

We used DataStax Cassandra to implement our Analysis, a distributed NoSQL database that delivers continuous availability, high performance, and linear scalability and supports Cassandra Query Language (CQL) with which we queried our data. DataStax has many advantages like we used DataStax Bulk Loader for Apache Cassandra to load our CSV into the table and DataStax Studio to query, explore, and visualize CQL with ease. We also used DataStax search for filtering CQL queries.

2.3 Methods incorporated

1. The first step to querying using Cassandra is to create a Keyspace. A keyspace is a container for our application data (like Schema in RDBMS). The keyspace requires that the replication strategy and replication factor be specified — the number of node data must be distributed as replicas.
2. In the next step, we created a table within the keyspace.
3. For inserting data into the table, dsbulkloader (provided by DSE) was installed.

4. Next, we created a search index on ‘jobtitle’ column to retrieve the count of rows that have ‘Data Analyst’ text.
5. The table was then queried using the ‘solr_query’ keyword which uses Apache Solr for full-text search. It took just 0.064s to search ~2500 rows.
6. Querying data was done with Datastax Studio. Results from the table were queried using the ‘JSON’ keyword to achieve JSON format.

3 EXPERIMENTAL DISCUSSION

In this section, we describe our projects implementation strategy, training and testing processes that we have implemented in this work. The approach we took was taking data cleaning it and experimenting with useful information that can be retrieved from the collected data. Then taking the queries and visualizing them to better understand the data related jobs.

3.1 Docker Images for DataStax Server and Datastax studio:

We installed the DataStax server (Cassandra distributor) on Docker.

Why Docker?

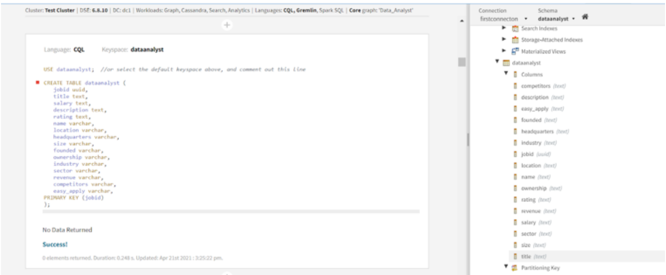
Docker Engine is mostly a Linux layer (Regardless of Windows/Mac). It helps in the containerization of the Apps. The Containers are portable. They contain the code, dependencies, config, process, networking which provides ease of use.

We used DataStax Docker images to create DataStax Enterprise (DSE) 6.8 server and DataStax Studio 6.8 containers following the document [5].

DataStax Enterprise is the always-on database designed to effortlessly build and scale our apps, integrating graph, search, analytics, administration, developer tooling, and monitoring into a single unified platform.

DataStax Studio is an interactive tool for CQL (Cassandra Query Language) and DataStax Enterprise Graph.DataStax Studio enables developers to query, explore, and visualize CQL and DSE Graph data with ease.

DataStax Cassandra: Cassandra delivers continuous availability (zero downtime), high performance, and linear scalability that modern applications require.



Cluster: Test Cluster | DSE: 6.8.10 | DC: dc1 | Workloads: Graph, Cassandra, Search, Analytics | Languages: CQL, Gremlin, Spark SQL | Core graph: 'Data_Analyst'

Language: CQL Keyspace: dataanalyst

```
USE dataanalyst;  //or select the default keyspace above, and comment out this line

CREATE TABLE dataanalyst (
   jobid uuid,
   title text,
   salary text,
   description text,
   rating text,
   name varchar,
   location varchar,
   headquarters varchar,
   size varchar,
   founded varchar,
   ownership varchar,
   industry varchar,
   sector varchar,
   revenue varchar,
   competitors varchar,
   easy_apply varchar,
PRIMARY KEY (jobid)
);
```

No Data Returned

Success!

0 elements returned. Duration: 0.248 s. Updated: Apr 21st 2021 : 3:25:22 pm.

Figure 3. This shows how table was created in Cassandra DB.

Dsbulkloader: DataStax Bulk Loader for Apache Cassandra® is open-source software (OSS). Use the product to load and unload CSV or JSON data in and out of supported databases.

Apache Solr: It’s major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document handling.

Knowi: Knowi natively integrates with Cassandra and DataStax for analytics and visualization.

3.2 Data Cleaning and preparation:

When we initially downloaded the dataset, we had an extra column that was only counting the rows. It was not necessary nor would it be helpful so decided to remove it. This gave a total of 16 columns and 12 columns have missing values. 'Easy Apply' and 'Competitors' have a maximum number of missing values (> 50%). In the Data Cleansing part, we removed the unnamed column.

3.3 Loading Data and Visualization:

We loaded a Data Analyst jobs dataset from Kaggle having more than 2000 job listing for data analyst positions, with the following features:

- Salary Estimate
- Job Title
- Location
- Company Rating
- Job Description
- Company Name
- Headquarters
- Type of Ownership
- Industry
- Sector within Industry
- Size of Company
- Revenue
- Founded Year
- Competitors
- EasyApply

into Cassandra. We used DataStax Studio for querying the loaded data.

4 VISUALIZATION

Dataset from Kaggle was loaded sequentially into Dastax Cassandra and then Knowi via the Cloul9Agent which connects Knowi and data source. Additional data cleaning was required to visualize map-related information in Knowi. Knowi does not support our dataset’s geo-information in ‘Location’ but supports only ‘City in USA’ or geographic coordinate system (longitude, latitude). So, we converted the feature, ‘Location’, to latitude and longitude and split it’s geoinformation into ‘city’ and ‘state’ with python. Then, we could visualize the dataset like 'Top 5 state of highest average salary' or 'The number of companies of top 5 states'

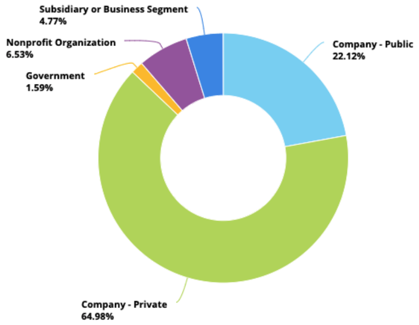


Figure 4. This shows the company type where the jobs are separated by.

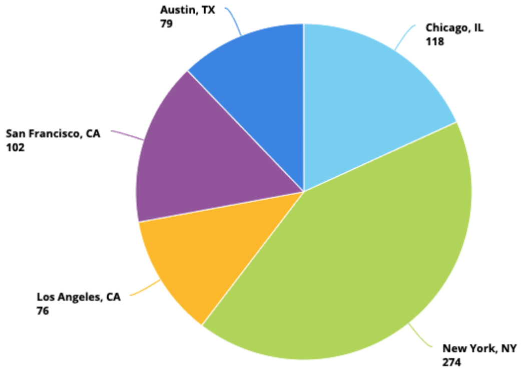


Figure 5. This shows the top states wher the most jobs from all the 3 categories are.

5 CONTRIBUTIONS

1. *Shahraiz Qureshi* – Team Lead, Data Collection, Data Cleansing, Managing scrum and agile
2. *Jennifer Kwok* – Data Modelling, Data Processing
3. *Gangotri Biswal* – Optimizing queries using DSE search and Apache solr, Data Visualization, CQL queries
4. *Chitra Priyaa* – Data loading and Querying data - CQLqueries - JSON, Geospatial data.
5. *Seung min Yoo - Data Visualization with Knowi, CQLqueries*
6. *Palak Shah - Designing workflow, data processing,* Apache solr queries, *Data Visualization,*

6 CONCLUSIONS

We found that the maximum number of Data Analysts jobs are available in New York followed by Chicago, San Francisco, Austin and Los Angeles. Data Analysts Glassdoor salary package ranges between 27K and 132K depending on the Company and individual's capacity.

The top 2 industries (IT Services and Staffing & Outsourcing) make up 34% of all Data Analyst jobs with a non-null Industry label.

References

[1]https://www.kaggle.com/andrewmvd/data-analyst-jobs

[2]https://docs.datastax.com/en/dse/5.1/cql/

[3]https://docs.datastax.com/en/studio/6.8/

[4]https://www.knowi.com/docs/cassandra.html

[5]https://docs.datastax.com/en/docker/doc/docker/dockerQuickStart.html

[6]https://github.com/picklesueat/data_jobs_data

[7] https://www.youtube.com/watch?v=fFkszsiVRw4

[8]https://www.youtube.com/watch?v=s1xc1HVsRk0&list=PLalrWAGybpB-L1PGA-NfFu2uiWHEsdscD