AskReddit Dataset Analysis

PostgreSQL vs Neo4J

SOEN 363 - Data Systems for Software Engineers

Info about the dataset

The dataset is called AskReddit that is contained within the dataset are 189,565 questions posted by reddit users and a total of 5,940,827 answers to these questions. split into multiple files.

- **reddit_questions.csv** contains 189,565 questions along with a unique id, a timestamp the number of upvotes received. delimiter is semicolon. **(27.4 MB)**
- **reddit_answers.csv** contains 5,566,660 answers to the questions along with the corresponding question id and upvotes. delimiter is semicolon. **(972.8 MB)**
- **reddit*answers*long.csv** contains all answers in reddit_answers.csv plus additional answers greater than 1000 characters long. delimiter is semicolon. **(1.37 MB)**

- Our main focus was performing queries on the **following scenario**:

Questions and answers in relation to Amazon and its CEO.

Data Loading

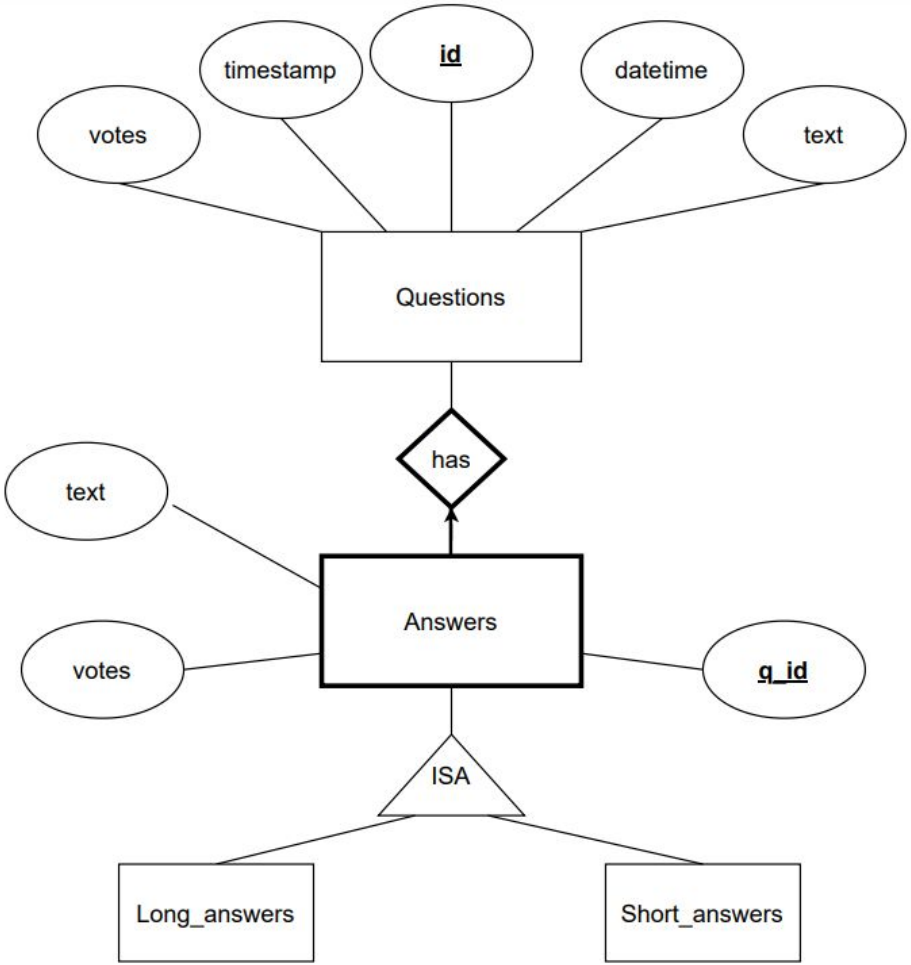
PostgreSQL	Neo4J
● 3 Schemas has been created to host the 3 tables. ● No issues were found while importing the CSV files	● The double quotations needed to be replaced by single quotation. ● Some tables had issues in delimiters, so cypher was raising errors. For instance, the reddit_answers_long was not fully imported due to null objects that has been indicated with extra delimiters.

11 queries

Query Number	Idea/Purpose behind the query
1	Find the questions that mention the word Amazon. Report the Id and the question's text.
2	Find all the answers whether long or short that have the word Amazon and that answer can be located in both datasets "shortAnswers and LongAnswers". Report their question's Id and the answer's text.
3	Find the questions whose answer has the word Amazon. Report the question and their count.
4	Find all the questions that have more than 100 answers and report question id and their count ascending.
5	Find all the questions that have answers in long and short answers, report the question and the count of the answers.
6	Find all the questions that have the word "Jeff" the owner of Amazon". Report the question with its time of posting.
7	Find the question with less than 3 votes that mentions the word "Amazon". Report the questions along with the votes.
8	Find the questions that have "Jeff Bezos" the owner of Amazon. Report the time at which it was posted and the date as well.
9	Find the List the questions of less than a 10 characters in them
10	Find the question that has the word "CEO" and was posted during the weekend. Report the datetime,timestamp and its corresponding number of votes .
11	Find for every question, the answer that has the maximum number of votes. Report the question text and the number of votes for the highest answer.


PostgreSQL

ER Diagram


id
timestamp
datetime
votes
text
Questions
has
text
Answers
votes
q_id
ISA
Long_answers
Short_answers

DDL Statements

```sql
create table reddit_questions
(
    id        varchar(50) not null primary key,
    text      text,
    votes     integer,
    timestamp numeric,
    datetime  date
);

create table reddit_answers
(
    q_id  varchar(50),
    text  text,
    votes numeric,
    primary key (q_id),
    foreign key (q_id) references reddit_questions(id) ON DELETE CASCADE ON UPDATE CASCADE
);


create table reddit_answers_long
(
    q_id  varchar(50),
    text  text,
    votes numeric,
    primary key (q_id),
    foreign key (q_id) references reddit_questions(id) ON DELETE CASCADE ON UPDATE CASCADE
);
```

Query 1: *Find the questions that mention word Amazon. Report the Id and the question text.*

```
SELECT id,text FROM reddit_questions
WHERE text LIKE '%amazon%' OR text LIKE '%Amazon%'
```

Services

postgres@localhost
- reddit_questions
  - reddit_questions
- console 662 ms
  - console 662 ms
- reddit_answers
- reddit_answers_long

Output | postgres.public.reddit_questions

39 rows | Tx: Auto | DDL

	id	text
1	b898b/	With all this talk about using reddit to make money, …
2	fsbih5	So, let's pretend something. The year ""2020"" is on …
3	itssby	How do you write a good Amazon review?
4	hn5vp/	Has anyone ever gotten a book published on the Kindle…
5	d6ocg/	Amazon.com phishing scam?
6	g98919	You've been enlarged into a 30ft Amazonian battle mag…
7	146e0o	What is something I can buy on Amazon for $20 or less…
8	vdp5b/	Buckaroo Banzai, Amazon Women on the Moon, and The Ke…
9	do0a74	What is the funniest Amazon review you've ever read?
10	hmuua/	Reddit I need your help, theres a group of people on …

Version Control | TODO | Problems | Database Changes | Services

39 rows retrieved starting from 1 in 561 ms (execution: 455 ms, fetching: 106 ms)

Query 2: *Find all the answers whether long or short that has word Amazon and that answer can be located in both datasets "shortAnswers and LongAnswers". Report their question's Id and the answer's text.*

```sql
SELECT q_id,text FROM reddit_answers_long
WHERE text LIKE '%amazon%' OR text LIKE '%Amazon%'
```

postgres@localhost
- reddit_questions
  - reddit_questions
- console 2 s 695 ms
  - console 2 s 695 ms
- reddit_answers
- reddit_answers_long

Output | postgres.public.reddit_answers_long

1-500 of 501+ | Tx: Auto | DDL

	q_id	text
1	cz6fi1	Leasing/Payment Plans- It artificially raises the cos…
2	hbq6gt	My favorite thing is when apple, Gmail, Amazon, or ot…
3	dpwp80	"The "problem with your amazon account" emails. Also,…
4	dpwp80	""You have won an Amazon Gift Card" "
5	dpwp80	At this moment in time I keep getting a lot of spam t…
6	i93mj8	I've had two really good experiences. I took my car …
7	hmzvu6	"Those spam bot things that say they're from a compan…
8	fg9d10	"Many "Amazon's Choice" products have been manipulate…
9	fg9d10	Amazon
10	ej6ewl	Amazon Flex, Deliveroo, Fiverr, ForAScore, Freelancer…

Version Control | TODO | Problems | Database Changes | Services

Query 3: *Find the questions whose answer has word Amazon. Report the question and their count.*

```sql
SELECT Q.text,COUNT(*) FROM reddit_answers_long L,reddit_questions Q
WHERE L.text LIKE '%amazon%' OR L.text LIKE '%Amazon%' AND  L.q_id=Q.id
GROUP BY Q.text
```

Services: All Services

- postgres@localhost
  - reddit_answers_long
  - console 1 m 8 s 596 ms
    - console 1 m 8 s
  - reddit_questions 133 ms
  - reddit_answers

Output | Result 20

1-500 of 501+

	text	count
1	"" 'You too will marry a boy I choose,' said Mrs Rupa M…	2019
2	"" For atheists, if there was no Godly surveillance wha…	2019
3	"" i remember when i had a right knee and left knee, no…	2019
4	"" in few years, dating will be limited to the upper 30…	2019
5	"" The difference in the temperature along the differen…	2019
6	""... is there even any other way to... hey, were you e…	2019
7	""/s"" is a largely successful written indicator for sa…	2019
8	""[Serious]"" People who have used papaya seeds as a ma…	2019
9	""[Serious]"" What drug, besides liquor and weed, can y…	2019
10	""10, 9, 8, 7, 6, 5, 4, 3, 2, 1! Happy New Year!"" You …	2019

Bookmarks

Query 4: *Find all the questions that have more than 100 answers and report their count ascending.*

```
SELECT Q.text,COUNT(*) FROM reddit_answers_long L,reddit_questions Q
WHERE  L.q_id=Q.id
GROUP BY Q.text
HAVING COUNT(*)>100
ORDER BY COUNT(*) ASC
```

Services

postgres@localhost
- reddit_questions
  - reddit_questions
- console 26 s 487 ms
  - console 26 s
- reddit_answers
- reddit_answers_long

Output | Result 11

1-500 of 501+

	text	count
1	What is an essential, not-so-obvious skill in life?	101
2	If, one day, we cure all diseases and aging, what wou…	101
3	At what moment did you realize it was time to quit yo…	101
4	What is your favourite song that has the word ""song"…	101
5	How would you cause the most chaos with the ability t…	101
6	Wealthy redditors, what are some services or products…	101
7	Wealthier redditors, how did you get your wealth and …	101
8	What series was consistently excellent from start to …	101
9	What's the nicest way to tell someone to F off?	101
10	You are the ruler of Heck, the lesser version of Hell…	101

Version Control | TODO | Problems | Database Changes | Services

Query 5: *Find all the questions that have answers in long and short answers, report the question and the count of the answers.*

```
SELECT Q.text,COUNT(*) FROM reddit_answers_long L,reddit_questions Q
WHERE  L.q_id=Q.id
GROUP BY Q.text
```

Services

postgres@localhost
- reddit_questions
  - reddit_questions
- console 36 s 5 ms
  - console 36 s
- reddit_answers
- reddit_answers_long

Output | Result 12

1-500 of 501+

	text	count
1	"" 'You too will marry a boy I choose,' said Mrs Rupa…	6
2	"" For atheists, if there was no Godly surveillance w…	24
3	"" i remember when i had a right knee and left knee, …	5
4	"" in few years, dating will be limited to the upper …	26
5	"" The difference in the temperature along the differ…	4
6	""... is there even any other way to... hey, were you…	6
7	""/s"" is a largely successful written indicator for …	25
8	""[Serious]"" People who have used papaya seeds as a …	4
9	""[Serious]"" What drug, besides liquor and weed, can…	9
10	""10, 9, 8, 7, 6, 5, 4, 3, 2, 1! Happy New Year!"" Yo…	17

Version Control | TODO | Problems | Database Changes | Services

Bookmarks

Query 6: *Find all the questions that have the word "Jeff" the owner of Amazon". Report the question with its time for posting*

```
-- Query #6
-- Find all the questions that have the word "Jeff" the owner of Amazon".
-- Report the question with its time of posting
select text as Question, timestamp
from reddit_questions
where text like '%Jeff%' or
      text like '%jeff%';
```

of votes . | Report the question ...r the highest answer. | Report the question ...r the highest answer. 2 | Report the question ..

103 rows

	question	timestamp
1	Coming from NA we hear about Jeffrey Dahmer, Ted Bundy, etc... What a…	1601420242
2	What would be the first thing you did if you traded bank accounts wit…	1575478174
3	What superhero or supervillain should Jeff Goldblum play?	1466965262
4	If Jeff Bezos exploded into his net worth of hay pennies, skyrim styl…	1595462893
5	Colin Peterson of MN and Jeff Drew of NJ are two Democrats who voted …	1572537750
6	What movie would you remake with Nicholas Cage as the lead role and J…	1548542333
7	We all know Jeff Bezos is currently the richest person on paper (net …	1599497685

Query 7: *Find all the question with less than 3 votes that mentions the word "Amazon"*

```
select text as question, votes as Num_votes
from reddit_questions
where (text like '%Amazon%' or
      text like '%amazon%') and votes < 3;
```

f votes . | Report the question ...r the highest answer. | Report the question ...r the highest answer. 2 | Report the quest

12 rows

	question	num_votes
1	So, let's pretend something. The year ""2020"" is on Amazon. Now that …	2
2	How do you write a good Amazon review?	2
3	who made a buisnes on Amazone ?	0
4	What is the most useful/interesting item you have bought from Amazon r…	2
5	What kind of product does amazon push you to try?	2
6	There were huge fires in the USA and in Congo, even bigger than the on…	2
7	Has anyone ever gotten a book published on the Kindle through Amazon?	2
8	What are some ways that amazon alexa has made your life easier?	2
9	What happened with the Amazone bushfires and what was the aftermath?	2
10	New Discovery Big underground river flows below Amazon	0
11	If you were in a competition and given a $150 amazon gift card what wo…	2
12	How do you feel about Amazons CEO Jeff Bezos being the first trilliona…	1

Query 8: *Find all the question that have "Jeff Bezos" the owner of Amazon. Report the time at which it was posted and the date as well*

```
-- Query #8
-- Find the questions that have "Jeff Bezos" the owner of Amazon.
-- Report the time at which it was posted and the date as well.

select text as question, timestamp, datetime
from reddit_questions
where text like '%Jeff Bezos%';
```

of votes . | Report the question ...r the highest answer. | Report the question ...r the highest answer. 2 | Report the question ...r the highest answer. 3 | Rep

48 rows

	question	timestamp	datetime
1	What would be the first thing you did if you traded bank accounts with…	1575478174	Wed Dec 4 16:49:34 2019 UTC
2	If Jeff Bezos exploded into his net worth of hay pennies, skyrim style…	1595462893	Thu Jul 23 00:08:13 2020 UTC
3	We all know Jeff Bezos is currently the richest person on paper (net w…	1599497685	Mon Sep 7 16:54:45 2020 UTC
4	Redditors who've thought of eating the rich, what would Jeff Bezos thi…	1597791273	Tue Aug 18 22:54:33 2020 UTC
5	You have the entire net worth of Jeff Bezos in your account for 1 hour…	1596332591	Sun Aug 2 01:43:11 2020 UTC
6	What is the first thing you'd do/buy if Jeff Bezos handed all his mone…	1601295720	Mon Sep 28 12:22:00 2020 UTC
7	How did people like Mark Zuckerberg, Jack Dorsey, Jeff Bezos and so on…	1603954641	Thu Oct 29 06:57:21 2020 UTC
8	Beyond all comprehension, you somehow kidnap Jeff Bezos. With his vast…	1596355765	Sun Aug 2 08:09:25 2020 UTC
9	Jeff Bezos decides to use his wealth to try and go the full Lex Luthor…	1573766143	Thu Nov 14 21:15:43 2019 UTC
10	Is Jeff Bezos the IRL equivalent to Lex Luthor or Mr. Krabs and why?	1585167298	Wed Mar 25 20:14:58 2020 UTC

Query 9: *Find the list the questions of less than a 10 characters in them*

```
-- Query #9
-- Find the List the questions of less than a 10 characters in them

select *
from reddit_questions
where length(text) < 10;



-- Query #10
-- Find the question that has the word "CEO" and was posted during the weekend.
-- Report the datetime,timestamp and its corresponding number of votes .
```

tes . | Report the question ...r the highest answer. | Report the question ...r the highest answer. 2 | Report the questio

11 rows | Tx: Auto | DDL

id	text	votes	timestamp	datetime
ced10/	Bad Names	10	1276378622	Sat Jun 12 21:37:02 2010 UTC
eh0wj/	Today	62	1291643779	Mon Dec 6 13:56:19 2010 UTC
e8rux/	Want $20?	178	1290192211	Fri Nov 19 18:43:31 2010 UTC
ps8o1/	First	0	1329396735	Thu Feb 16 12:52:15 2012 UTC
q2rnx/	TRAVEL	0	1330019834	Thu Feb 23 17:57:14 2012 UTC
13lzvd	SJ Cycles	0	1353567644	Thu Nov 22 07:00:44 2012 UTC
wo5cr/	VF	0	1342480690	Mon Jul 16 23:18:10 2012 UTC
q02ns/	Snoo toy?	5	1329868202	Tue Feb 21 23:50:02 2012 UTC
doxbj/	10/10/10	16	1286616085	Sat Oct 9 09:21:25 2010 UTC
mfb9u/	Help!	2	1321504528	Thu Nov 17 04:35:28 2011 UTC
cbxyi/	DAE	0	1275805525	Sun Jun 6 06:25:25 2010 UTC

Query 10: *Find the questions that have the word "CEO" and was posted during the weekend" Report the datetime, timestamp and its corresponding number of votes*

```
-- Query #10
-- Find the question that has the word "CEO" and was posted during the weekend.
-- Report the datetime,timestamp and its corresponding number of votes .

select text as question, votes, timestamp, datetime
from reddit_questions
where text like '%CEO%' and (datetime like '%Sat%' or datetime like '%Sun%');
```

Report the question ...r the highest answer. | Report the question ...r the highest answer. 2 | Report the question ...r the highest ar

21 rows

question	votes	timestamp	dat
ou are now the CEO of Vault-Tec. Whats the most evil,inhumane, or hum…	4078	1485001185	Sat J
EOs and Production Managers of any meat producing companies of Reddit…	2	1582432856	Sun F
ongrats! You're the new CEO of Hell! What new ideas do you bring to t…	178	1571591123	Sun O
hich CEOs definitely do not use his/her own product?	5187	1481982295	Sat D
o the people with neck and forearm tattoos, did it really affect your…	1	1605478431	Sun N
Serious] CEOs, government employees, military, and IC redditors: What…	28	1549834022	Sun F
f Superman were real, which present day CEO would be his Lex Luthor…	17	1305477092	Sun M

Query 11: Find for every question, the answer that has the maximum number of votes. Report the question text and the number of votes for the highest answer.

```
-- Query #11
-- Find for every question, the answer that has the maximum number of votes.
-- Report the question text and the number of votes for the highest answer.

select rq.id, rq.text, max(ra.votes)
from reddit_questions as rq, reddit_answers_long as ra
where rq.id = ra.q_id
group by rq.id, rq.text
order by max(ra.votes) desc ;
```

est answer. 3 | Report the question ...r the highest answer. 4 | Report the datetime,...ing number of votes . 2 | Report the question ...r the highe

181,465 rows

	id	text	max
1	fkzaca	What is something that has aged well?	99398
2	9yc7op	What's the rudest thing a guest has ever done in your home?	90303
3	a0a4cd	What's the most amazing thing about the universe?	86042
4	d0jjc2	The 2010's decade will be over in 4 months. What do you think people …	85936
5	aqf3bi	You are offered $1,000,000 USD if you can hide a pair of car keys fro…	85693
6	bvdaci	What's classy if you're rich but trashy if you're poor?	85568

Using Indexing Techniques

Based on the proposed queries, indexing some tables on specific attributes made the queries run faster which made the queries more efficient:

```
CREATE INDEX question_id_index for reddit_questions (id)
CREATE INDEX short_answers_index for reddit_answers (q_id)
CREATE INDEX long_answers_index for reddit_answers_long (q_id,votes)
```

After creating an indexes on the tables, we found time reduction in running the queries:

Query 1: *Find the questions that mention word Amazon. Report the Id and the question text.*

It went from 455 ms to 80 ms with 72% time reduction

```sql
SELECT id,text FROM reddit_questions
WHERE text LIKE '%amazon%' OR text LIKE '%Amazon%'
```

Services: All Services

Output | postgres.public.reddit_questions

39 rows | Tx: Auto | DDL

	id	text
1	b898b/	With all this talk about using reddit to make money, how do you …
2	fsbih5	So, let's pretend something. The year ""2020"" is on Amazon. Now…
3	itssby	How do you write a good Amazon review?
4	hn5vp/	Has anyone ever gotten a book published on the Kindle through Am…
5	d6ocg/	Amazon.com phishing scam?
6	g98919	You've been enlarged into a 30ft Amazonian battle mage, the sun …
7	146e0o	What is something I can buy on Amazon for $20 or less that could…
8	vdp5b/	Buckaroo Banzai, Amazon Women on the Moon, and The Kentucky Frie…
9	do0a74	What is the funniest Amazon review you've ever read?
10	hmuua/	Reddit I need your help, theres a group of people on the Amazon …

Bookmarks

Version Control | TODO | Problems | Database Changes | Services

39 rows retrieved starting from 1 in 98 ms (execution: 80 ms, fetching: 18 ms)

Query 2: *Find all the answers whether long or short that has word Amazon and that answer can be located in both datasets "shortAnswers and LongAnswers". Report their question's Id and the answer's text.*

It From 2sec and 665 ms to 388 ms with 85% time reduction

```
SELECT q_id,text FROM reddit_answers_long
WHERE text LIKE '%amazon%' OR text LIKE '%Amazon%'
```

Services: All Services

Output | postgres.public.reddit_answers_long

1-500 of 501+ | Tx: Auto | DDL

	q_id	text
1	2bdwqj	About 99% of our products are available through Amazon for cheap…
2	fkzzn9	Amazon giving unpaid sick leave while demanding overtime and the…
3	fkzzn9	Amazon/WholeFoods. They're doing unpaid sick without an actual d…
4	8fs0ix	"I work at a GNC and I can safely vouch that a large percentage …
5	agviaf	Amazon. You have to sift through pages of garbage with company n…
6	agviaf	"Amazon, I know its not "products" but their customer service an…
7	agviaf	Amazon and their fake or counterfeit products
8	4uve7j	"Former employee for a "Computer Refurbishing" company that sell…
9	becepq	"Sears, Roebuck and Company, colloquially known as "Sears" - The…
10	becepq	Yelp, it used to be reviews but now their extortionist practices…

Version Control | TODO | Problems | Database Changes | Services

500 rows retrieved starting from 1 in 414 ms (execution: 388 ms, fetching: 26 ms)

Query 3: *Find the questions whose answer has word Amazon. Report the question and their count.*

IT went From 4 minutes and 41 seconds to 1 minute and 8 seconds with 75% time reduction

```
SELECT Q.text,COUNT(*) FROM reddit_answers_long L,reddit_questions Q
WHERE L.text LIKE '%amazon%' OR L.text LIKE '%Amazon%' AND L.q_id=Q.id
GROUP BY Q.text
```

postgres@localhost
- reddit_questions
  - reddit_questions
- console 4 m 41 s 699 ms
  - console 4 m 41 s
- reddit_answers
- reddit_answers_long

Output | Result 1

1-500 of 501+

	text	count
1	"" 'You too will marry a boy I choose,' said Mrs Rupa M…	2019
2	"" For atheists, if there was no Godly surveillance wha…	2019
3	"" i remember when i had a right knee and left knee, no…	2019
4	"" in few years, dating will be limited to the upper 30…	2019
5	"" The difference in the temperature along the differen…	2019
6	""... is there even any other way to... hey, were you e…	2019
7	""/s"" is a largely successful written indicator for sa…	2019
8	""[Serious]"" People who have used papaya seeds as a ma…	2019
9	""[Serious]"" What drug, besides liquor and weed, can y…	2019
10	""10, 9, 8, 7, 6, 5, 4, 3, 2, 1! Happy New Year!"" You …	2019

Version Control | TODO | Problems | Database Changes | Services

	Before Indexing	After Indexing	Time Reduction
Query #1	455ms	80ms	82%
Query #2	2 sec 665ms	388ms	85%
Query #3	4 min 41 sec	1 minute 8 sec	75%
Query #4	26 sec	8 sec	69%
Query #5	36 sec	9 sec	75%
Query #6	176 ms	154 ms	13%
Query #7	128 ms	105 ms	18%
Query #8	156 ms	153 ms	1.923%
Query #9	139 ms	126 ms	9%
Query #10	158 ms	84 ms	47%
Query #11	42 sec	6 sec	86%


neo4j

NoSQL Databases

- For this part we have decided to go with Graph Database Neo4J
- We chose to work on an online community dataset about AskReddit Questions and answers.
- Our main focus was performing queries on the following scenario:
  - Questions and answers in relation to Amazon and its CEO.

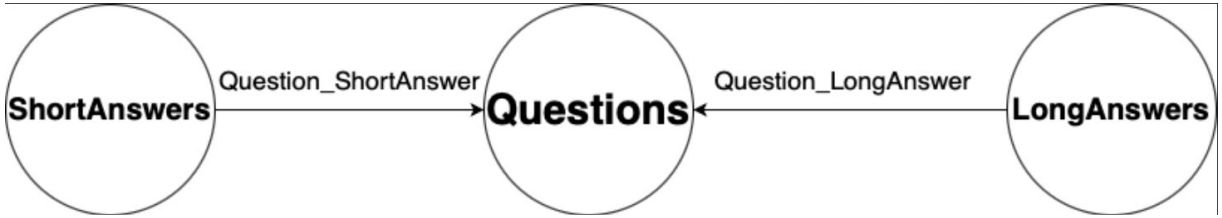
Creating Tables

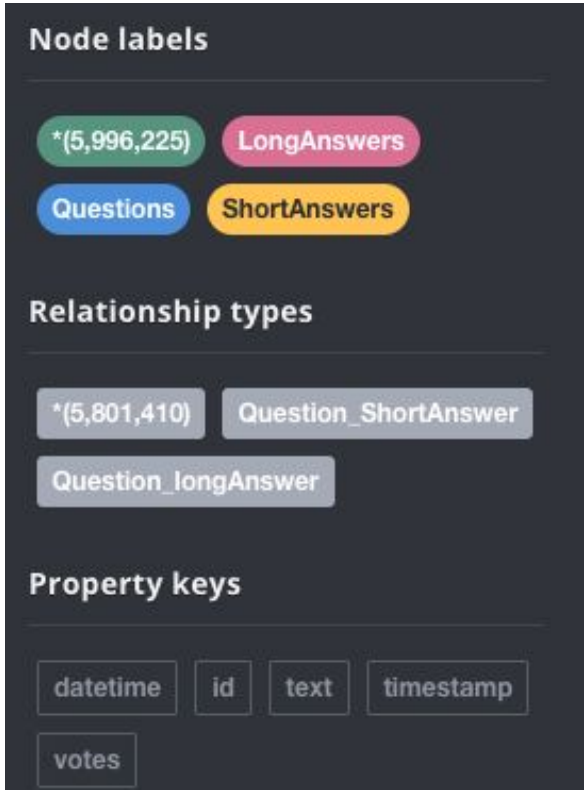
```
load csv with headers from "file:///reddit_questions_cleaned.csv" as row
FIELDTERMINATOR ';'
create (q:Questions) set q.id=row.id, q.text=row.text,
q.votes=toInteger(row.votes), q.timestamp=row.timestamp,
q.datetime=row.datetime

CALL apoc.periodic.iterate('CALL apoc.load.csv("reddit_answers_cleaned.csv" ,
{sep: \';\'}) yield map as row return row','CREATE (sa:ShortAnswers) set
sa.id=row.q_id, sa.text=row.text, sa.votes=toInteger(row.votes)',
{batchSize:10000, iterateList:true, parallel:true})


CALL apoc.periodic.iterate('CALL
apoc.load.csv("reddit_answers_long_cleaned.csv" , {sep: \';\'}) yield map as
row return row','CREATE (la:LongAnswers) set la.id=row.q_id, la.text=row.text,
la.votes=toInteger(row.votes)', {batchSize:100000, iterateList:true,
parallel:true})
```

Creating Relations

```
CALL apoc.periodic.iterate("
MATCH (q: Questions)
MATCH (sa: ShortAnswers)
WHERE q.id = sa.id
RETURN q, sa",
"CREATE (q)-[r:Question_ShortAnswer]->(sa)",
{}) YIELD batches, total, errorMessages
RETURN batches, total, errorMessages

CALL apoc.periodic.iterate("
MATCH (q: Questions)
MATCH (la: LongAnswers)
WHERE q.id = la.id
RETURN q, la",
"CREATE (q)-[r:Question_longAnswer]->(la)",
{}) YIELD batches, total, errorMessages
RETURN batches, total, errorMessages
```

Data Model


ShortAnswers
Question_ShortAnswer
Questions
Question_LongAnswer
LongAnswers


Node labels
*(5,996,225)
LongAnswers
Questions
ShortAnswers
Relationship types
*(5,801,410)
Question_ShortAnswer
Question_longAnswer
Property keys
datetime
id
text
timestamp
votes

Consistency Vs. Availability

Neo4j is dedicated to providing basic availability as a NoSQL system.

- Strong consistency is granted due to a system that uses a page-cache and a transaction log to record changes to the graph.
- When reading local writes, Neo4j manages **causal consistency** and guarantees **eventual consistency**, at some point.
- Causal consistency ensures that all processes agree on the causal sequence of operations. for instance, When one process is the result of another, for instance, two operations are said to be **causally related**.

Query 1: *Find the questions that mention word Amazon. Report the Id and the question text.*

```
Match (a:Questions)
WHERE a.text contains "Amazon"
    RETURN a.id, a.text
```

Table | Text | Code

	a.id	a.text
1	"b898b/"	"With all this talk about using reddit to make money, how do you feel about submissions linking to Amazon with embedded affiliate
2	"fsbih5"	"'So, let's pretend something. The year ''''2020'''' is on Amazon. Now that the 90-day trial is over, what's your review?'"
3	"itssby"	"How do you write a good Amazon review?"
4	"hn5vp/"	"Has anyone ever gotten a book published on the Kindle through Amazon?"
5	"d6ocg/"	"Amazon.com phishing scam?"
6	"g98919"	"You've been enlarged into a 30ft Amazonian battle mage, the sun is setting over the distant horizon and the sound of war echoes

Started streaming 35 records after 2 ms and completed after 1156 ms.

Query 2: *Find all the answers whether long or short that has word Amazon and that answer can be located in both datasets "shortAnswers and LongAnswers". Report their question's Id and the answer's text.*

```
Match (a:LongAnswers)
Match (b:ShortAnswers)
WHERE a.text contains "Amazon" AND a.id=b.id
    RETURN Distinct a.id, a.text
```

Table | Text | Code

	a.id	a.text
1	"hvbvpz"	"An air compressor to air up your tires. Got one off Amazon for 30 bucks."
2	"ea803g"	"A 500 pieces puzzle, not much, but I would love to make a puzzle and clear my mind. Edit: freaking wow... Thank you so much guys for the av
3	"fgi5bd"	"Star Trek The Next Generation is available on Amazon for free for all Prime members. All 7 seasons. Edit. My wife as I started watching it and
4	"9799el"	"''''Buy Amazon gift cards in cash Publish several ''books'' on Amazon Buy my books with gift cards Money laundered I cannot find the story, b
5	"aiv6l6"	"''''I work as a Creative Director. I have a lot of great clients, unfortunately with a few shitty managers from their side. They usually go with the n
6	"cy82ym"	"Surprised no one has mentioned the Panama Papers. They've been out for a while, but there's been little coverage of them. TLDR of them: Lo

Started streaming 297 records after 1 ms and completed after 16563 ms.

Query 3: *Find the questions whose answer has word Amazon. Report the question and their count.*

```
Match (a:ShortAnswers)<-[]-(q:Questions)
WHERE a.text contains "Amazon"
    RETURN Distinct q.text, count(a) as count order by count desc
```

Table | Text | Code

	q.text	count
2	"'If Jeff Bezos woke up tomorrow and said ''''lets fuck up the worlds economy'''' what would be the best way he could do it?'"	21
3	"Without exceding 100$, which monthly subscriptions (like Netflix) should everyone subscribe to?"	17
4	"A question for Whole Foods employees: How do you feel the culture at Whole Foods has changed since the Bezos take over? If it hasn't changed, have new procedures been implemented?"	17
5	"If Jeff Bezos died and left his fortune to whoever completed a some sort of challenge, what challenge would be most appropriate?"	16
6	"Overlooked authors of Reddit, what is your book about and where can we read it?"	15
7	"What do you think the scariest corporation is?"	14
8		

Started streaming 2903 records after 222 ms and completed after 253 ms, displaying first 1000 rows.

Query 4: *Find all the questions that have more than 100 answers and report their count ascending.*

```
Match (a:ShortAnswers)←[]-(q:Questions)
WITH q, count(a) AS count
WHERE count>100
    RETURN Distinct q.text, count order by count desc
```

	q.text	count
1	"What bot accounts on reddit should people know about? (self.AskReddit)"	276
2	"The nearest green thing kills you. What do you die from?"	273
3	"The world ended. You are dead. The credits to your life roll. What song would you want to play during?"	271
4	"If your life was a book, what is the name of the current chapter?"	268
5	"If you were to die today, what would your headstone say if it had to be brutally honest?"	267
6	"If you were arrested, and your username was the charge, what crime did you commit?"	267
7		

Started streaming 17639 records after 3 ms and completed after 16 ms, displaying first 1000 rows.

Query 5: *Find all the questions that have answers in long and short answers, report the question and the count of the answers.*

```
Match (a:ShortAnswers)<-[]-(q:Questions)
Match (b:LongAnswers)<-[]-(q:Questions)
WITH q,(count(a)+count(b)) AS allAnswers
    RETURN Distinct q.text, allAnswers order by allAnswers desc
```

	q.text	allAnswers
1	"What bot accounts on reddit should people know about? (self.AskReddit)"	152352
2	"You are going to be murdered. The last person you searched on Google is your weapon to defend yourself, what did you get?"	137288
3	"You now have to fight the last thing you Googled. The object to your left is your weapon, and the object to your right is your defense. The first number you can think of is how much HP you have. How screwed are you?"	124002
4	"What's your favorite yelled movie line?"	121032
5	"If 2020 had a final boss you had to fight at the end of the year, what weapon (real or fictional) would you want to use?"	119072
6	"'You wake up tomorrow to a Doc taking off brain monitoring sensors from your head. He hands you a hefty cheque and says '''Thank you for taking part in our hyper-real sensory experiment, project 2020'''. The actual date is December 1, 2019. You leave the building. What is the first thing you do?'"	118098
7		

Started streaming 4097 records after 2 ms and completed after 3 ms, displaying first 1000 rows.

Query 6: *Find all the questions that has the word "Jeff" the owner of Amazon. Report the question with its time of posting*

```
neo4j$ MATCH (n:Questions) WHERE n.text contains "Jeff" Return
       DISTINCT n.text, n.timestamp
```

Table | Text | Code

	n.text
1	"Coming from NA we hear about Jeffrey Dahmer, Ted Bundy, etc... What are the European/African/Asian continent equival
2	"What would be the first thing you did if you traded bank accounts with Jeff Bezos?"
3	"What superhero or supervillain should Jeff Goldblum play?"
4	"If Jeff Bezos exploded into his net worth of hay pennies, skyrim style, How big of an impact would the resulting explosion
5	"Colin Peterson of MN and Jeff Drew of NJ are two Democrats who voted no today on the impeachment rules. Anyone kn
6	"We all know Jeff Bezos is currently the richest person on paper (net worth), but who is likely to be the the person with the

Started streaming 101 records after 2 ms and completed after 697 ms.

Query 7: Find the question with less than 3 votes that mentions the word "Amazon". Report the questions along with the votes

```
neo4j$ MATCH (n:Questions) with n, count(n.votes) AS count WHERE count
       <3 RETURN n.text, count order by count desc
```

Table | Text | Code

	n.text
995	"Programmers of reddit, what was your first completed project?"
996	"Can we find out what happened to this guy? Real or fake? The creepiest thing to ever happen on Reddit, ever."
997	"Is there any way to permanently change how hair grows from curly to straight?"
998	"'Husbands of ''''bridezillas'''', how did you handle your wife's attitude on your wedding day and did it change your image
999	"If I were to see a nuclear mushroom cloud in the distance, what are the first steps I should take to survive?"
1,000	"Where are some unusual places you've been?"

Started streaming 947825 records after 3 ms and completed after 6 ms, displaying first 1000 rows.

Query 8: Find the questions that have "Jeff Bezos" the owner of Amazon. Report the time at which it was posted and the date as well.

```
neo4j$ MATCH (n:Questions) WHERE n.text contains "Jeff
       Bezos" RETURN n.text,n.timestamp, n.datetime
```

Table | Text | Code

	n.text
1	"What would be the first thing you did if you traded bank accounts with Jeff Bezos?"
2	"If Jeff Bezos exploded into his net worth of hay pennies, skyrim style, How big of an impact w
3	"We all know Jeff Bezos is currently the richest person on paper (net worth), but who is likely to
4	"Redditors who've thought of eating the rich, what would Jeff Bezos thighs go good with?"
5	"What is the first thing you'd do/buy if Jeff Bezos handed all his money to you?"
6	"You have the entire net worth of Jeff Bezos in your account for 1 hour, with 2 minutes notice. V

Started streaming 240 records after 2 ms and completed after 810 ms.

Query 9: Find the List the questions of less than a 10 characters in them

```
neo4j$ MATCH (n:Questions) WHERE size(n.text) < 10
       RETURN n.text
```

Table | Text | Code

	n.text
1	"Bad Names"
2	"Today"
3	"Want $20?"
4	"First"
5	"TRAVEL"
6	"SJ Cycles"
7	

Started streaming 55 records after 3 ms and completed after 789 ms.

Query 10: Find the question that has the word "CEO" and was posted during the weekend. Report the datetime,timestamp and its corresponding number of votes .

```
neo4j$ MATCH (n:Questions) WHERE n.text contains "CEO" AND
       (n.datetime contains "Sat" OR n.datetime contains "Sun")
       return n.text,n.datetime,n.timestamp, n.votes
```

Table | Text | Code

	n.datetime	n.timestamp	n.votes
	"Sat Jan 21 12:19:45 2017 UTC"	"1485001185.0"	4078
	"Sun Feb 23 04:40:56 2020 UTC"	"1582432856.0"	2
	"Sun Oct 20 17:05:23 2019 UTC"	"1571591123.0"	178
usual episode like?'"	"Sat Feb 1 00:23:09 2020 UTC"	"1580516589.0"	5
gs when hiring or does it simply depend on the type of tat?"	"Sun Nov 15 22:13:51 2020 UTC"	"1605478431.0"	1
	"Sun Nov 30 19:09:50 2014 UTC"	"1417374590.0"	1553

Started streaming 105 records after 1 ms and completed after 517 ms.

Query 11: Find for every question, the long answer that has the maximum number of votes. Report the question text and the number of votes for the highest answer.

```
Match (b:LongAnswers)<-[]-(q:Questions)
WITH q, max(b.votes) AS maxVotes
    RETURN Distinct q.text, maxVotes order by maxVotes desc
```

	q.text	maxVotes
1	"What is something that has aged well?"	99398
2	"What's the most amazing thing about the universe?"	86042
3	"The 2010's decade will be over in 4 months. What do you think people will remember this decade for?"	85936
4	"You are offered $1,000,000 USD if you can hide a pair of car keys from the entirety of the FBI force for 7 days. Where do you hide the keys?"	85693
5	"What's classy if you're rich but trashy if you're poor?"	85568
6	"What is the greatest comeback to a insult you've ever heard?"	85107
7		

Started streaming 4101 records after 22 ms and completed after 25 ms, displaying first 1000 rows.

Using Indexing Techniques

Based on the proposed queries, indexing some tables on specific attributes made the queries run faster which made the queries more efficient:

```
create Index question_id for (q:Questions) on (q.id);
create Index ShortAnswer_question_id for (sa:ShortAnswers) on (sa.id);
create Index LongAnswers_votes for (la:LongAnswers) on (la.votes);
```

After creating an indexes on the tables, we found time reduction in running the queries:

Query 1 after Indexing: *Find the questions that mention the word Amazon. Report the Id and the question text.*

We noticed that It went from 1156 ms to 309 ms with **73.269%** time reduction

```
Match (a:Questions)
WHERE a.text contains "Amazon"
    RETURN a.id, a.text
```

Table | Text | Code

	a.id	a.text
1	"b898b/"	"With all this talk about using reddit to make money, how do you feel about submissions linking to Amazon with embedded affiliate
2	"fsbih5"	"'So, let's pretend something. The year ''''2020'''' is on Amazon. Now that the 90-day trial is over, what's your review?'"
3	"itssby"	"How do you write a good Amazon review?"
4	"hn5vp/"	"Has anyone ever gotten a book published on the Kindle through Amazon?"
5	"d6ocg/"	"Amazon.com phishing scam?"
6	"g98919"	"You've been enlarged into a 30ft Amazonian battle mage, the sun is setting over the distant horizon and the sound of war echoes

Started streaming 35 records after 1 ms and completed after 309 ms.

Query 2 after Indexing: *Find all the answers whether long or short that has word Amazon and that answer can be located in both datasets "shortAnswers and LongAnswers". Report their question's Id and the answer's text.*

We noticed that It went from 16563 ms to 7250 ms with **56.228%** time reduction.

```
Match (a:LongAnswers)
Match (b:ShortAnswers)
WHERE a.text contains "Amazon" AND a.id=b.id
    RETURN Distinct a.id, a.text
```

Table | Text | Code

	a.id	a.text
1	"hvbvpz"	"An air compressor to air up your tires. Got one off Amazon for 30 bucks."
2	"ea803g"	"A 500 pieces puzzle, not much, but I would love to make a puzzle and clear my mind. Edit: freaking wow... Thank you so much g
3	"fgi5bd"	"Star Trek The Next Generation is available on Amazon for free for all Prime members. All 7 seasons. Edit. My wife as I started wa
4	"9799el"	"'''Buy Amazon gift cards in cash Publish several ''books'' on Amazon Buy my books with gift cards Money laundered I cannot fi
5	"aiv6l6"	"'''I work as a Creative Director. I have a lot of great clients, unfortunately with a few shitty managers from their side. They usually
6	"cy82ym"	"Surprised no one has mentioned the Panama Papers. They've been out for a while, but there's been little coverage of them. TLD

Started streaming 297 records after 2 ms and completed after 7250 ms.

Query 3 after Indexing: *Find the questions whose answer has word Amazon. Report the question and their count.*

We noticed that It went from 253 ms to 4 ms with **98.4%** time reduction

```
Match (a:ShortAnswers)<-[]-(q:Questions)
WHERE a.text contains "Amazon"
    RETURN Distinct q.text, count(a) as count order by count desc
```

Table | Text | Code

	q.text
1	"'When online shopping, what's the most dubious/weird thing you've had recommended to you in the ''''Customers who bought XXXXX also bo
2	"'If Jeff Bezos woke up tomorrow and said ''''lets fuck up the worlds economy'''' what would be the best way he could do it?'"
3	"Without exceding 100$, which monthly subscriptions (like Netflix) should everyone subscribe to?"
4	"A question for Whole Foods employees: How do you feel the culture at Whole Foods has changed since the Bezos take over? If it hasn't chang
5	"If Jeff Bezos died and left his fortune to whoever completed a some sort of challenge, what challenge would be most appropriate?"
6	"Overlooked authors of Reddit, what is your book about and where can we read it?"

Started streaming 2903 records after 2 ms and completed after 4 ms, displaying first 1000 rows.

	Before Indexing	After Indexing	Time Reduction
Query #1	1156 ms	309ms	73.269%
Query #2	16563 ms	7250 ms	56.228%
Query #3	253 ms	4 ms	98.4 %
Query #4	16 ms	7 ms	56.25%
Query #5	3 ms	2 ms	33.33%
Query #6	697 ms	478 ms	31.4203 %
Query #7	1000 ms	1000 ms	0%
Query #8	810 ms	475 ms	41.358%
Query #9	789 ms	504 ms	36.1216 %
Query #10	517 ms	486 ms	5.99%
Query #11	25 ms	4 ms	84%

THANK YOU

Any Questions?