Cal Poly Pomona

Safa Alasady

Exploring Big Data and Its Applications

5/11/2024

CIS 3050.05

Professor Ahmed Azam

Spring 2024

Table of Contents

Statement of Academic Honesty

My name is: Safa Alasady, I declare that, except where fully referenced no aspect of this project has been copied from any other source. I understand that any act of Academic Dishonesty such as plagiarism or collusion may result in serious offense and punishments. I promise not to lie about my academic work, to cheat, or to steal the words or ideas of others, nor will I help fellow students to violate the Code of Academic Honesty.

Name: Safa Alasady Date: 5/11/2024

Signature: 

Introduction

Today, the application and usage of data is extremely important, especially in the era of technology. Organizations go through large amounts of data and these amounts of data grow every day. Since Big Data has become very valuable to big organizations, different ways of handling data have been developed like Hadoop and Hive. Hadoop and Hive enable more scalability and durability of data. These approaches to big data have been used by Google and Facebook. Issues emerge from organizations that try to collect large amounts of data because it sometimes is not able to fit in with the relational model. Because big data has resulted in unstructured data, non-relational databases are being used to fit this data in the databases. Non-relational databases have become a competition to the usually accepted traditional relational databases. Organizations like Amazon use non-relational databases for their vast amount of data. Because large amounts of data result in bigger databases, being able to retrieve data needs to be quick and efficient. Sometimes data can be lost and unstructured and this data can be very valuable. Through data analytics, organizations can learn a lot about trends and information about customers. Not only can organizations learn about trends and customer information, but they can also make good decisions, predict patterns, and solve any problems that can occur. There are also different types of data analytics like predictive, prescriptive data analytics, etc. Organizations can store and run their large databases on a cloud rather than a computer to save storage and space. This project will lean into the more practical side of data and databases. It will allow students to learn more about the applications of data in a real-life environment.

Q1:

A new trend which is called Big Data is being introduced. Big Data comes from large amounts of data. Because of how large this data can be, it can become lost and unstructured, and it does not fit into existing models. Because of issues surrounding large databases, many organizations like Google use non-relational solutions that are called NoSQL databases. The Hadoop Distributed File System is also a new way to process unstructured data. A way to store and retrieve data made by business orders is to use SQL and relational databases. One limitation of handling large amounts of unstructured data is that SQL only supports unstructured data through text or string. NoSQL uses Hadoop technology, and it can be involved in already existing or new data warehouses. NoSQL has been transformed into non-relational, distributed, and scalable databases. There are some standard criteria for NoSQL databases to be able to fit in with OLAP and other systems. Data Vault is a model that uses both the third normal form and the star schema. Anchor modeling is a database modeling technique that surrounds the idea that data warehouse environments are always changing and small changes in the models should only occur because of these big changes in the outside world. MapReduce was developed by Google. MapReduce is a programming model, and it makes the allocation of tasks easier across multiple nodes. Hadoop utilizes a lot of data, however ETL is better for smaller amounts of data since it reduces the time needed to build data warehouses. Hadoop is an open-source programming framework. HDFS is the Hadoop distributed file system and it is a scalable distributed file system that provides quick access to data. There are downsides to Hadoop. Because of these downsides, MapReduce programs must be written for small and simple inquiries. Hive was created to develop an interface that can deal with unstructured data in Hadoop with tables, columns, rows, functions, and methods similar to SQL. HiveQL is a language that uses features like joint, grouping, and aggregation just like SQL.

Q2:

The main purpose of SQL and relational databases is to enable organizations to be able to store and retrieve data created by business orders happening daily like orders and salaries. SQL and relational databases can execute queries using this data and based on specific filters, it can also generate reports. There are problems when it comes to using SQL along with large amounts of data. SQL supports unstructured text search but only through text which has led to the development of NoSQL models.

Q3:

The difference between relational databases and non-relational databases is when it comes to large amounts of data or Big Data. Large amounts of data can be hard to process and the solution to that is a NoSQL model, which again, is a non-relational database. Relational databases are good to use for daily business operations, like invoices, to get queries using this data and create reports. A non-relational database's main task is to make sure that it can retrieve and store data fast. There are also criteria or standards for non-relational databases, which are column-oriented, document-oriented, key-value database, and graph databases. The criteria for relational databases are atomicity, consistency between columns, and uniqueness.

Q4:

The difference between Hadoop and ETL tools is that Hadoop uses large amounts of data while ETL is better for small amounts of data since it saves time for data warehouse development time. Sometimes Hadoop provides quick access to high-quality data without having to use ETL tools. Also, ETL tools have a lot of tools that are already available like cleaning while in Hadoop, one would have to manually write these tasks. However there is a disadvantage for ETL, ETL tools are not scalable and can face issues when trying to go through large amounts of data. Because of this issue, it is better to use ETL for smaller amounts of data, and then for large amounts of data, Hadoop would be the better option. Through custom programs, Hadoop can use ETL, but it can be at times not useful and sometimes can't be used instead of traditional ETL tools.

Q5:

Hive query language is a language that is very similar to SQL. It uses the same features as JOIN, cartesian products, grouping, aggregation, creating tables, SELECT command, etc. Because these are also used in SQL, anyone who understands SQL can easily start writing queries. SQL has problems with adding new columns or rows to a table and to do so, a person would have to rewrite the table, but with HiveQL data is retrieved periodically making it easier to load data into a new table. HiveQL allows for more complex algorithms.

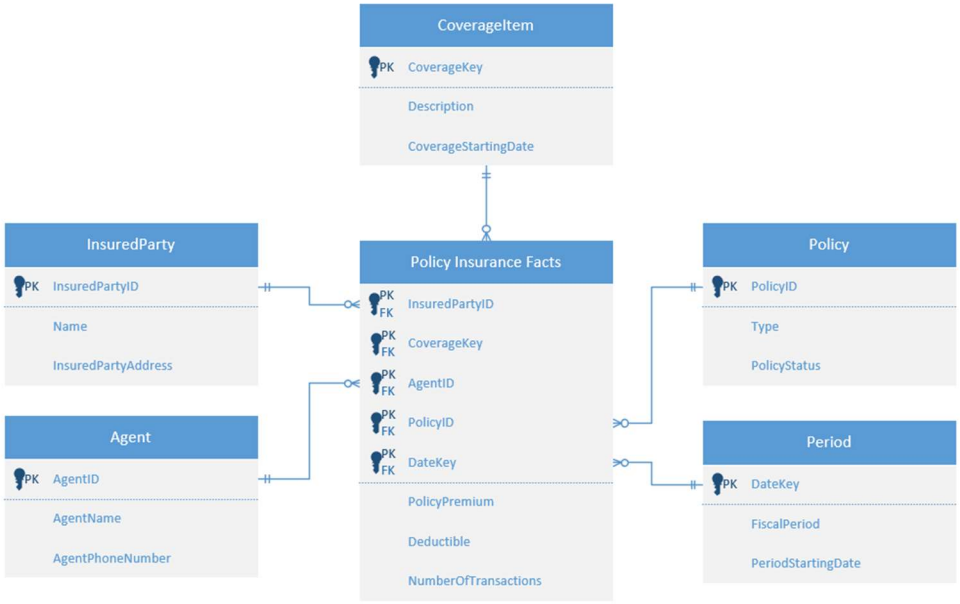
Q6:

A star schema is "A simple database design that consists of a group of tables that describe the dimensions of the business arranged logically around a huge central table that contains all the accumulated facts and figures of the business", according to Modern Database Management by Jeffrey Hoffer. There are two variations of a star schema, multiple fact tables and factless fact tables. There is also another version of a star schema which is the snowflake schema. The snowflake schema is an expanded version of a star schema where the dimension outer tables are normalized into more separate tables. The star schema below is a standard star schema with provided information used from question 9-39. I also added different attributes for creativity like the CoverageStartingData attribute under the CoverageItem table and the InsuredPartyAddress attribute under the InsuredParty table. I also found the total rows and size using the assumptions of a fiscal period being one month, 5 years of historical data in the data mart, and the assumption that there will be a 5 percent change each month.

Assumptions:

- The length of the fiscal period is one month.
- Five years of historical data in the data mart.
- There will be approximately 5 percent of the policies that experience some type of change each month.

A)



B) Total rows = 0.05 (change) * 1,000,000 approximate number of policies * 10 covered items per policy * 2 insured parties * 5 years of historical data in data mart * 12 months per year

= 60,000,000 rows

C)

Total size = 60,000,000 total rows * 8 fields * 10 bytes

= 4,800,000,000 bytes or 4.80 gigabytes

Q7:

According to the book Modern Database Management, an index speeds up random and sequential access to base table data. The command to use index is part of the data definition language. There are two commands for using an index, CREATE INDEX and DROP INDEX. The CREATE INDEX command makes searching through columns easy to query. The DROP INDEX command will of course delete/remove an index. Indexes also do not affect data. It can take a long time to look for information without an index since it will have to look through every row. An index is good to use when it comes to sorting, joins, and constraints, and when your database is large. Indexes also help to remove duplicate rows. There are also some drawbacks to indexes. Indexes take up more space in the database. Indexes slow down the INSERT, UPDATE, and DELETE commands.

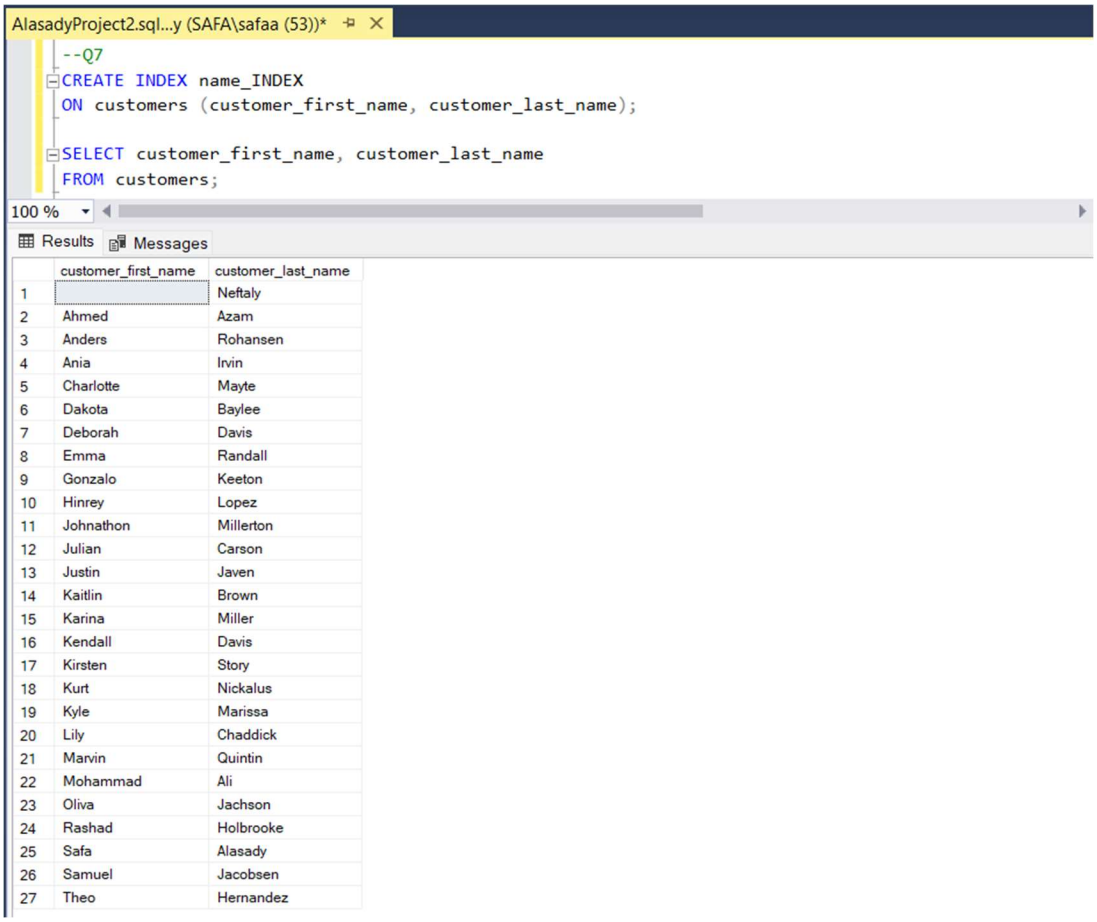
```
--Q7

CREATE INDEX name_INDEX

ON customers (customer_first_name, customer_last_name);


SELECT customer_first_name, customer_last_name

FROM customers;
```



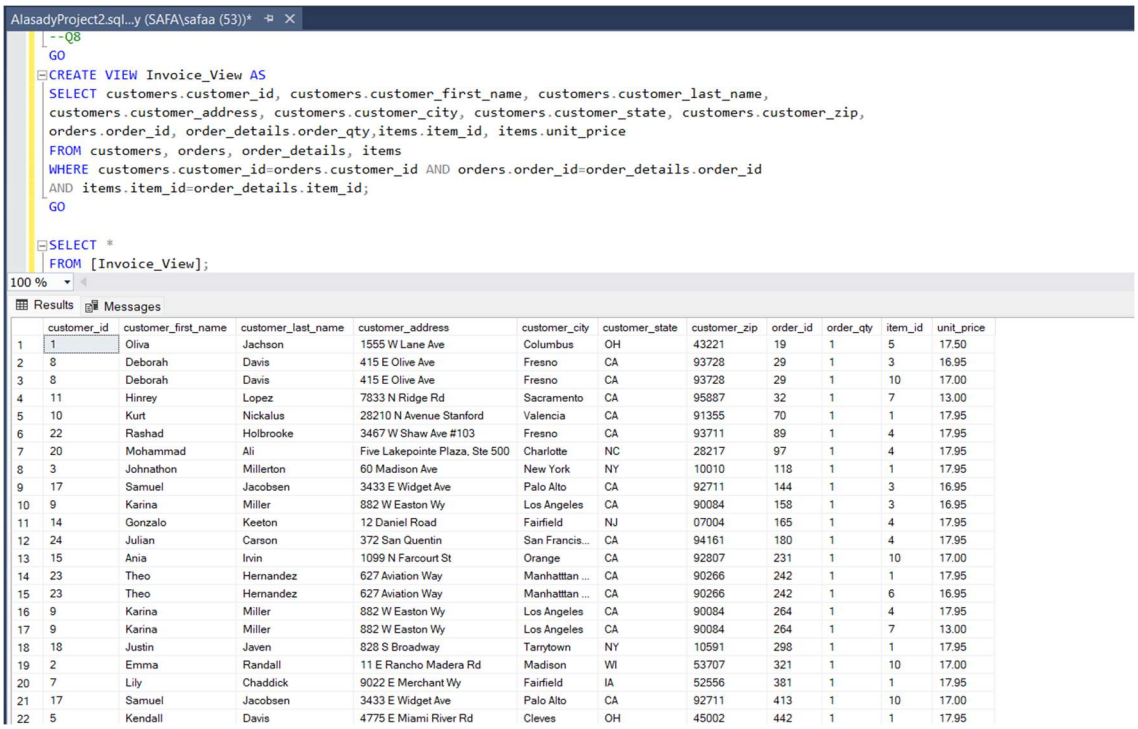
	customer_first_name	customer_last_name
1		Neftaly
2	Ahmed	Azam
3	Anders	Rohansen
4	Ania	Irvin
5	Charlotte	Mayte
6	Dakota	Baylee
7	Deborah	Davis
8	Emma	Randall
9	Gonzalo	Keeton
10	Hinrey	Lopez
11	Johnathon	Millerton
12	Julian	Carson
13	Justin	Javen
14	Kaitlin	Brown
15	Karina	Miller
16	Kendall	Davis
17	Kirsten	Story
18	Kurt	Nickalus
19	Kyle	Marissa
20	Lily	Chaddick
21	Marvin	Quintin
22	Mohammad	Ali
23	Oliva	Jachson
24	Rashad	Holbrooke
25	Safa	Alasady
26	Samuel	Jacobsen
27	Theo	Hernandez

Q8:

An invoice requires details about the customer like their name, address, and customer ID. It also requires details about the order and items that are within the order like order ID, order quantity, item ID, and unit price. In specific, the data elements necessary to create an invoice for a customer would be customer_id, customer_first_name, customer_last_name, customer_address, order_id, order_qty, item_id, and unit_price.

```
--Q8
GO
CREATE VIEW Invoice_View AS
SELECT customers.customer_id, customers.customer_first_name,
customers.customer_last_name,
customers.customer_address, customers.customer_city, customers.customer_state,
customers.customer_zip,
orders.order_id, order_details.order_qty,items.item_id, items.unit_price
FROM customers, orders, order_details, items
WHERE customers.customer_id=orders.customer_id AND
orders.order_id=order_details.order_id
AND items.item_id=order_details.item_id;
GO

SELECT *
FROM [Invoice_View];
```



AlasadyProject2.sql...y (SAFA\safaa (53))*

```
--Q8
GO
CREATE VIEW Invoice_View AS
SELECT customers.customer_id, customers.customer_first_name, customers.customer_last_name,
customers.customer_address, customers.customer_city, customers.customer_state, customers.customer_zip,
orders.order_id, order_details.order_qty,items.item_id, items.unit_price
FROM customers, orders, order_details, items
WHERE customers.customer_id=orders.customer_id AND orders.order_id=order_details.order_id
AND items.item_id=order_details.item_id;
GO

SELECT *
FROM [Invoice_View];
```

100 %

Results | Messages

	customer_id	customer_first_name	customer_last_name	customer_address	customer_city	customer_state	customer_zip	order_id	order_qty	item_id	unit_price
1	1	Oliva	Jachson	1555 W Lane Ave	Columbus	OH	43221	19	1	5	17.50
2	8	Deborah	Davis	415 E Olive Ave	Fresno	CA	93728	29	1	3	16.95
3	8	Deborah	Davis	415 E Olive Ave	Fresno	CA	93728	29	1	10	17.00
4	11	Hinrey	Lopez	7833 N Ridge Rd	Sacramento	CA	95887	32	1	7	13.00
5	10	Kurt	Nickalus	28210 N Avenue Stanford	Valencia	CA	91355	70	1	1	17.95
6	22	Rashad	Holbrooke	3467 W Shaw Ave #103	Fresno	CA	93711	89	1	4	17.95
7	20	Mohammad	Ali	Five Lakepointe Plaza, Ste 500	Charlotte	NC	28217	97	1	4	17.95
8	3	Johnathon	Millerton	60 Madison Ave	New York	NY	10010	118	1	1	17.95
9	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	144	1	3	16.95
10	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	158	1	3	16.95
11	14	Gonzalo	Keeton	12 Daniel Road	Fairfield	NJ	07004	165	1	4	17.95
12	24	Julian	Carson	372 San Quentin	San Francis...	CA	94161	180	1	4	17.95
13	15	Ania	Irvin	1099 N Farcourt St	Orange	CA	92807	231	1	10	17.00
14	23	Theo	Hernandez	627 Aviation Way	Manhatttan ...	CA	90266	242	1	1	17.95
15	23	Theo	Hernandez	627 Aviation Way	Manhatttan ...	CA	90266	242	1	6	16.95
16	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	264	1	4	17.95
17	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	264	1	7	13.00
18	18	Justin	Javen	828 S Broadway	Tarrytown	NY	10591	298	1	1	17.95
19	2	Emma	Randall	11 E Rancho Madera Rd	Madison	WI	53707	321	1	10	17.00
20	7	Lily	Chaddick	9022 E Merchant Wy	Fairfield	IA	52556	381	1	1	17.95
21	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	413	1	10	17.00
22	5	Kendall	Davis	4775 E Miami River Rd	Cleves	OH	45002	442	1	1	17.95

23	1	Oliva	Jachson	1555 W Lane Ave	Columbus	OH	43221	479	2	1	17.95
24	1	Oliva	Jachson	1555 W Lane Ave	Columbus	OH	43221	479	1	4	17.95
25	16	Dakota	Baylee	1033 N Sycamore Ave.	Los Angeles	CA	90038	491	1	6	16.95
26	3	Johnathon	Millerton	60 Madison Ave	New York	NY	10010	523	1	9	16.95
27	2	Emma	Randall	11 E Rancho Madera Rd	Madison	WI	53707	548	1	9	16.95
28	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	550	1	1	17.95
29	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	550	1	4	17.95
30	16	Dakota	Baylee	1033 N Sycamore Ave.	Los Angeles	CA	90038	601	1	5	17.50
31	16	Dakota	Baylee	1033 N Sycamore Ave.	Los Angeles	CA	90038	601	1	9	16.95
32	20	Mohammad	Ali	Five Lakepointe Plaza, Ste 500	Charlotte	NC	28217	607	1	3	16.95
33	20	Mohammad	Ali	Five Lakepointe Plaza, Ste 500	Charlotte	NC	28217	607	1	10	17.00
34	2	Emma	Randall	11 E Rancho Madera Rd	Madison	WI	53707	624	1	7	13.00
35	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	627	1	9	16.95
36	20	Mohammad	Ali	Five Lakepointe Plaza, Ste 500	Charlotte	NC	28217	630	1	5	17.50
37	20	Mohammad	Ali	Five Lakepointe Plaza, Ste 500	Charlotte	NC	28217	630	2	6	16.95
38	12	Anders	Rohansen	12345 E 67th Ave NW	Takoma Park	MD	24512	651	1	3	16.95
39	12	Anders	Rohansen	12345 E 67th Ave NW	Takoma Park	MD	24512	658	1	1	17.95
40	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	687	1	6	16.95
41	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	693	1	6	16.95
42	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	693	3	7	13.00
43	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	693	1	10	17.00
44	19	Kyle	Marissa	789 E Mercy Ave	Phoenix	AZ	85038	703	1	4	17.95
45	13		Neftaly	2508 W Shaw Ave	Fresno	CA	93711	778	1	1	17.95
46	13		Neftaly	2508 W Shaw Ave	Fresno	CA	93711	778	1	3	16.95
47	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	796	1	5	17.50
48	17	Samuel	Jacobsen	3433 E Widget Ave	Palo Alto	CA	92711	796	1	7	13.00
49	19	Kyle	Marissa	789 E Mercy Ave	Phoenix	AZ	85038	800	1	1	17.95
50	19	Kyle	Marissa	789 E Mercy Ave	Phoenix	AZ	85038	800	1	5	17.50
51	2	Emma	Randall	11 E Rancho Madera Rd	Madison	WI	53707	802	1	3	16.95
52	1	Oliva	Jachson	1555 W Lane Ave	Columbus	OH	43221	824	1	3	16.95
53	1	Oliva	Jachson	1555 W Lane Ave	Columbus	OH	43221	824	2	7	13.00
54	18	Justin	Javen	828 S Broadway	Tarrytown	NY	10591	827	1	6	16.95
55	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	829	1	1	17.95
56	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	829	1	5	17.50
57	9	Karina	Miller	882 W Easton Wy	Los Angeles	CA	90084	829	1	9	16.95

Q9:

A query optimizer in SQL is the process/part that is responsible for choosing the best way to execute queries with the least amount of time and resources. In *Modern Database Management*, "The query optimizer chooses the best plan based on statistics about each table, such as average row length and number of rows" (page 358). The query optimizer works in the background. SQL query is different from query optimizer. SQL query is the instructions you would give to the query optimizer to execute these instructions in the best way possible. SQL query allows users to retrieve and manipulate data while the query optimizer takes place after writing the SQL query. For example, a simple SQL query could be SELECT product name FROM products WHERE price > 10. After the query is written and executed by the user, the query optimizer would go through different ways and conclude which one saves time and resources to produce the results required by the user.

Q10:

It is stated in Modern Database Management that data analytics is "the systematic analysis and interpretation of data typically using mathematical, statistical, and computational tools- to improve our understanding of a real-world domain" (page 444). In other words, data analytics is the process of evaluating data to obtain information that will allow individuals to learn and draw conclusions to make decisions. Data analytics can improve business decisions because of the ability to make decisions based on insights or reports that could come from data analytics. Businesses could look at trends that stem from data analytics and make a knowledgeable decision. There are multiple types of data analytics, predictive, prescriptive, diagnostic, and descriptive. Predictive data analytics answers the question of what could happen. It uses statistics and computational methods against data from the past and present to understand what could happen in the future. An example of predictive data analytics is a business using data analytics of data regarding their customer's buying behavior to understand what other products they could produce to make a profit. Prescriptive data analytics could advise on possible

outcomes. Prescriptive data analytics uses the results of predictive data analytics to understand what a business should do to reach a desired conclusion. An example of prescriptive data analytics is when a business uses predictive data analytics to examine data surrounding customers to make certain actions like advertising products similar to a customer's past purchase. Diagnostic data analytics answers the question of why this event happened. Diagnostic data analytics looks at data to understand why a specific outcome has occurred. For example, a business could see that sales for a product are dropping and understand why they are dropping using diagnostic data analytics through patterns and trends. Descriptive data analytics answers the question of what has happened. Descriptive data analytics analyzes the past of data to gain an understanding of the current trends surrounding a topic of interest. For example, a retail business uses data analytics from the past to understand the current trends and patterns of fashion styles for shoes. Structured data usually comes from databases and data warehouses. Structured data is usually in the form of tables, relations, arrays, etc. The data types for structured data are numeric, character, and dates. Unstructured data comes from databases and other sources like multimedia files and emails. Unstructured data comes in the form of documents, e-mails, maps, etc. Semi-structured data comes from sources like social media. Semi-structured data comes in the form of XML, JSON files, etc.

Q11:

Cloud computing is "A model for provisioning and acquiring computing services on demand using centralized resources that are accessed either through the public Internet or a private network", stated Modern Database Management (page 373). There are four approaches to cloud computing, Infrastructure-as-a-Service (laaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS), and Database-as-a-Service (DBaaS). Infrastructure-as-a-Service is a cloud computing approach in which its services consist of hardware and systems software resources. Platform-as-a-Service (PaaS) consists of Infrastructure-as-a-Service resources which will allow developers to produce a high level of productivity with laaS. Software-as-a-Service (SaaS) consists of software and application solutions to focus on the needs of a noncompeting activity. Database-as-a-Service consists of the usage of a data management platform service. Cloud computing is not viable to use for my project since my database is too small and it would not make sense to use the cloud. It would be more beneficial for larger databases because these businesses with large databases, who can provide cloud services to their customers, can run their server and operating systems and can save more money by adopting cloud computing.

Analysis

Many of the topics like big data, non-relational databases, Hadoop, and more gave me in-depth learning and understanding. The article provided gave a closer look at not only big data, non-relational databases, and Hadoop but also at NoSQL databases, MapReduce, and Hive. Starting from the beginning, the purpose of SQL and relational databases was discussed with the discussion being that SQL and RDBMS enable organizations to be able to store and retrieve data created by business orders happening daily. The differences between relational and non-relational databases were also discussed. Big data was the difference when it came to using relational or non-relational databases. It was also analyzed that there are criteria for relational databases which are atomicity, consistency, and uniqueness. It was found that Hadoop is better for bigger databases while ETL tools should be used for smaller databases. Another thing that was explored was that HiveQL is very similar to SQL in the ways that it uses JOIN and SELECT commands, and many other similar features. The most difficult part of this project was creating the star schema and calculating the total size needed for question 9-39. It was difficult because I had to go through multiple drafts of the star schema. I made several errors, for example, when it came to the dimension tables and missing attributes for those tables. Besides these different topics and other topics that were analyzed in the project, I learned about the real-life application of big data and the solutions to solve the issues that come with big data and large databases. I learned that big organizations like Amazon, Facebook, and Google use non-relational databases for large amounts of data. In general, many important topics were discussed and questions relating to big data were answered.

Findings

I understood that data analytics existed, but I did now know that there were different detailed types of data analytics. I also did not understand the important applications of data analytics. This topic made me think about the uncomfortable idea that organizations are using my data to make decisions that could influence customers like me to make profitable decisions that could benefit a company. Besides that thought, it also enlightened me in learning how beneficial and invaluable data is to organizations. Continuing the issue that I had with creating the star schema, I was confused about how I would create the schema whether it would be with MS SQL or another program, but I did decide on doing it in Microsoft Visio for simplicity. In the end, I was able to create a star schema using the information provided. As part of my findings, I learned a slightly new topic, cloud computing. I know that there are clouds for many things like pictures, but I did not think databases would also be applicable.

Recommendations

My only recommendation is to include more information about the length of the answers to the questions. The instructions for this project are a lot clearer and straightforward compared to the first two projects, especially the first project. Throughout completing all three projects, I learned a lot in detail about the topics surrounding data. I enjoyed completing this project because it ties everything that I have learned from the previous projects together.

Lessons Learned

I learned many lessons like big data, non-relational databases, NoSQL, etc. I learned important topics like the many solutions for handling big data. Another lesson I learned was when to use relational and non-relational databases to handle data. Another thing I learned about the application of data analytics, especially figuring out what information an organization might need to be informed about things like customer behavior. Data analytics also proves to be extremely valuable for organizations to make smart decisions. I learned when cloud computing should be used, depending on the size of a database. Compared to reading the textbook, Modern Database Management, I learned the more practical side of databases and big data.

Conclusion

This project is a very good way to tie everything that I have learned this semester. The project was basically about writing a report using our past work and knowledge of projects 1 and 2. In a summary of my work for this project, I answered questions that talked about big data, Hadoop, Hive, types of databases, data analytics, cloud computing, etc. In a more in-depth description, I learned that big data comes with many problems but can prove to be of extreme importance to organizations. Big Data has created many different things like Hadoop and Hive. Big companies like Amazon, Google, and Facebook use non-relational databases to handle the large amounts of data they receive since relational databases cannot handle unstructured data that comes from Big Data. Allowing organizations to adapt to big data in this era of technology and data, non-relational databases provide scalability and can retrieve data quickly. Besides databases, data analytics is used to learn information about an organization's desired topic and make good decisions based on the knowledge they gain. There are several types of data analytics which include predictive, prescriptive, diagnostic, and descriptive analytics. Lastly, cloud computing is valuable depending on the size of a database. In the end, I thoroughly enjoyed completing this project as well as projects 1 and 2 and learning about data more in-depth.

References

- Tomislav Subic, et al. “Big Data in Data Warehouses.” ResearchGate, unknown, 2015, www.researchgate.net/publication/314643798_Big_Data_in_Data_Warehouses. Accessed 24 Apr. 2024.
- Hoffer, Jeffrey A, et al. Modern Database Management. 13th ed., Boston, Pearson Education, 2019.
- Simplilearn. “Index in SQL: Everything You Need to Know about It.” Simplilearn.com, Simplilearn, 18 May 2021, www.simplilearn.com/tutorials/sql-tutorial/index-in-sql. Accessed 30 Apr. 2024.
- Rouse, Margaret. “Query Optimizer.” Techopedia, 17 Dec. 2011, www.techopedia.com/definition/26224/query-optimizer. Accessed 30 Apr. 2024.
- GeeksforGeeks. “What Is Semi-Structured Data?” GeeksforGeeks, GeeksforGeeks, 12 Apr. 2019, www.geeksforgeeks.org/what-is-semi-structured-data/. Accessed 30 Apr. 2024.