Computer Society of India™ Since 1965

NASHIK CHAPTER: KNOWLEDGE SHARING SERIES: ARTICLE 658: 24 JULY 2023

Unlocking the power of data collaboration

Marcus Law

Using a collaborative approach to data can benefit organisations in several ways, driving success in a digital world

In 1978, Blockbuster Video founder David Cook formed a data services business.

Cook Data Services, which made software for oil and gas companies, went public in 1983, raising US$8.4m in an IPO.

But, six months later, the business had nosedived. Rather than sink the $8.4m into the now struggling company, he decided to put it toward a new business. Blockbuster was born.

With a vast network of retail locations across the United States and an extensive catalogue of movies, Blockbuster quickly became a symbol of the home entertainment industry. Despite its success, though, the company failed to anticipate the transformative power of data and the rapidly changing consumer preferences.

Netflix, on the other hand, embraced the digital revolution and recognised the potential of data to revolutionise the consumption of entertainment. What started as a subscription-based DVD-by-mail service in 1997 evolved into a streaming behemoth that disrupted the traditional rental model.

By leveraging data to understand consumer behaviour, preferences, and viewing patterns, Netflix could deliver personalised recommendations and a seamless streaming experience.

This transformation has been seen globally. In today's digital landscape, data has become the lifeblood of businesses across industries, fueling innovation, driving strategic decision-making, and enhancing customer experiences. As Fawad Qureshi, Industry Field CTO at data company Snowflake, explains, the company helps its clients make better use of data as well as aiding collaboration both in and outside of their organisation.

Collaboration and data cleanrooms

According to Snowflake, fast, streamlined access to relevant data is key to successful business outcomes. However, many organisations struggle to share data – internally between teams and externally with partners, vendors, and customers. Data collaboration has the power to improve the quality of analytics programs and even create new product offerings.

Sharing data while adhering to privacy regulations has always been challenging. But, by using distributed data clean rooms, it’s now possible to collaborate with data in a secure manner that aligns with privacy rules. Data clean rooms allow organisations to manage data effectively, deidentify it, and share it.

“I often describe a data clean room as an escrow account for data sharing,” Qureshi comments. “It's a middle account. Both parties share the data and control what you can do with that data.”

In 1982, computer scientist Andrew Yao proposed Yao's Millionaires' Problem: a secure multi-party computation problem which discusses two millionaires who are interested in knowing which of them is richer without revealing their actual wealth.

“The problem is there are two millionaires,” Qureshi comments. “Both of them want to know whether they have more money than the other, but they don't want to tell the other person how much money they have.”

According to Qureshi, the solution to this problem can be found in data clean rooms: a secure environment that allows multiple companies, or divisions of a single company, to bring data together for joint analysis under defined guidelines and restrictions.

“Data clean rooms are one of the best ways of doing data sharing without revealing sensitive and proprietary information, while also creating new business value for both sides.”

Changing how data collaboration works

Data collaboration is the process of gathering and sharing data from various sources. This process typically involves combining data sets from internal teams such as sales, marketing, and customer service and empowering domain experts to contribute their unique perspectives to inform insights. Data collaboration also takes the form of data-sharing partnerships or supplementing existing data with third-party data sets.

“Organisations have been doing data sharing and collaboration in many different ways,” says Qureshi, “but what we are doing is changing the way it is done.

“In the 1990s there was Blockbuster video. What did you do? You went on a weekend, got a cassette, took it home, and put it into the VCR.”

Think of this cassette as a copy of the data: the extract, transform, load (ETL) device.

“The VCR extracts the movie out of the cassette, transforms it, and projects it onto your screen,” Qureshi adds. “Later on, those VHS cassettes became CDs. You take the CD home and it has scratches; you put the CD into the CD drive – it doesn't work. Those scratches are data quality problems.

“Then came a company called Netflix. What they did was they said okay, instead of giving you copies of the data, the VHS will have a single copy of the movie. You bring your own compute, you bring your own bandwidth; you connect to the same copy of the movie and you can watch it at your leisure wherever you want. And whenever I want to end a movie, I can just say this movie is no longer available on Netflix.”

On top of that, users can gain insights into how the movies are being used. “What Snowflake is doing is the same Netflix-like experience for data sharing,” Qureshi says. “Instead of doing FTP and manual copies of the data, we have a single copy of the data. Consumers bring their own compute and we give them this DataFlix experience.”

As Qureshi explains, this data sharing experience has a number of benefits when compared to more traditional techniques.

“I have asked this question to many different organisations that are doing data sharing in traditional ways, can you tell me what is your most profitable, most valuable data feed? And they have no idea. Because when you create a copy of the data from FTP from here to here, you don't know what's happening on the other side. You don't get any usage information. But with Snowflake, you get metadata back on how many queries were run on your copy of the data, and how it has been consumed.”

Qureshi predicts that the future of data collaboration will see more and more growth: “We are going to get into more systematic data collaboration rather than the old Blockbuster-style of cassette sharing – because cassette sharing isn’t a scalable model and it doesn’t bring you any insights.”

Video of The Week

Explore some related information to above article at following link.

https://www.youtube.com/watch?v=iKrl6ldw-YE

https://www.youtube.com/watch?v=Mp7t3XY6MfY

https://www.youtube.com/watch?v=WCgWsXxMweU

https://www.youtube.com/watch?v=WwDy2t7O_iI

https://www.youtube.com/watch?v=i47N4eqrjtw

News of The Week

US Regulator Investigating ChatGPT Over Bad Content

OpenAI's release of ChatGPT last November stunned the world as it displayed the power of large language models (or LLM), a form of artificial intelligence known as generative AI that can churn out human-like content in just seconds.

The US Federal Trade Commission is investigating OpenAI to determine if its hugely popular ChatGPT app harms consumers by generating false information and whether its technology mishandles user data.

Microsoft-backed OpenAI was notified of the investigation in a 20-page questionnaire in which the company is asked to describe incidents in which users were falsely disparaged, and share any company efforts to ensure this does not happen again.

The Washington Post first reported the investigation by the US regulator.

OpenAI's release of ChatGPT last November stunned the world as it displayed the power of large language models (or LLM), a form of artificial intelligence known as generative AI that can churn out human-like content in just seconds.

Amid the marvel at the technology's capacities, reports came in that the models could also churn out offensive, false or just strange content, sometimes called "hallucinations".

FTC chair Lina Khan addressed a congressional committee hearing on Wednesday, and while she did not mention the investigation, she told lawmakers that her agency had concerns about ChatGPT's potentially libellous output.

"We've heard about reports where people's sensitive information is showing up in response to an inquiry from somebody else," Khan said.

"We've heard about libel, defamatory statements, flatly untrue things that are emerging. That's the type of fraud and deception that we are concerned about," she added.

The FTC's investigation is mainly focused on how this aspect could harm users, according to the questionnaire, but also delves into OpenAI's use of private data to build its world-leading model.

The company's GPT-4 is the bedrock technology behind its own ChatGPT as well scores of other programs from companies that pay a fee to OpenAI to access its model for their own uses.

An FTC probe does not necessarily bring further action and the regulator can close the case if it is satisfied by the target company's answer.

If the FTC perceives illegal or unsafe practices, it will demand remedial action and possibly launch a lawsuit.

OpenAI and the FTC did not respond to a request for comment.

E Toon

1. Feedback/Contribution of articles is appreciated at: nasikcsi@gmail.com

To know more about forthcoming programs and events, please do visit chapter website