A CASE STUDY ON

DUOLINGO

(based on big data analytics)

ABSTRACT

The ability to store and analyse vast volumes of data is critical for tailored and innovative user experiences in today's data-driven world. . This is proven by the successful use of analytics and big data to improve the services provided by Duolingo, an effective language learning website with more than 500 million users. Duolingo provide the world's largest language-learning database. To grasp distinct learning patterns., Duolingo gathers enormous volumes of user interaction data, which it examines in real time. Next, more sophisticated machine learning models enhance learning pathways and tailor lessons. In the digital learning space, Duolingo sets the standard by offering a personalized and interesting learning experience through constant analysis and adaptation to user data.

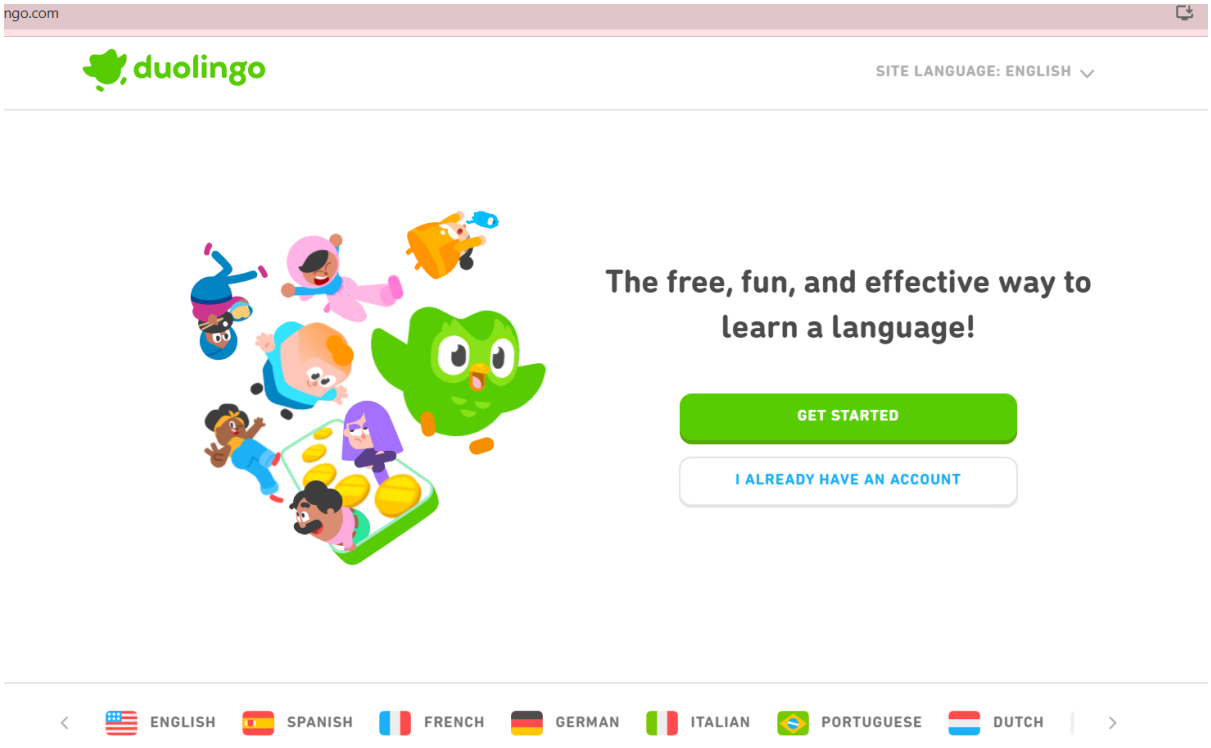

Fig 1.0

INTRODUCTION:

With more than 500 million users worldwide, Duolingo is a trailblazing language learning app that skillfully uses big data to streamline and customize the learning process. This case study delves extensively into the creative ways in which Duolingo

manages and analyzes its enormous user data by utilizing big data technologies and services. Its three main strategies are natural language processing to boost language understanding content, A/B testing for ongoing feature enhancement, and machine learning for personalized lesson recommendations. The core components of Duolingo's operational architecture are technologies like TensorFlow for machine learning, Google BigQuery for data warehousing, Apache Kafka for real-time data streaming, and Tableau for data visualization. Due to its strong infrastructure, Duolingo is able to handle and analyze large amounts of data quickly, which results in more engaging users and individualized learning experiences.

Fig 2.0

Big Data in Duolingo

Duolingo’s blend of accessible, engaging, and technologically advanced language learning makes it a standout in the field of digital education. Its commitment to leveraging big data, machine learning, and gamification has revolutionized how millions of people learn new languages. As Duolingo

continues to evolve, it remains at the forefront of making language learning both fun and universally accessible.

Big data signifies large amounts of organized and unstructured data gathered fast and in a variety of methods. It comprises data sets that are too large or complex to be handled efficiently by conventional information processing software. The 5 V's usually summarize the key elements of big data.
Duolingo's blend of accessible, engaging, and technologically advanced language learning makes it a standout in the field of digital education. Its commitment to leveraging big data, machine learning, and gamification has revolutionized how millions of people learn new languages. As Duolingo continues to evolve, it remains at the forefront of making language learning both fun and universally accessible.

Operational Workflow

1. Data Collection:
   Duolingo captures data concerning the languages a user is studying, the abilities and courses a user finished, and word quality (the words a user has shown to know). This also includes information about when an individual accesses Duolingo, the amount of lessons they done, and the in-app expenditures users make.

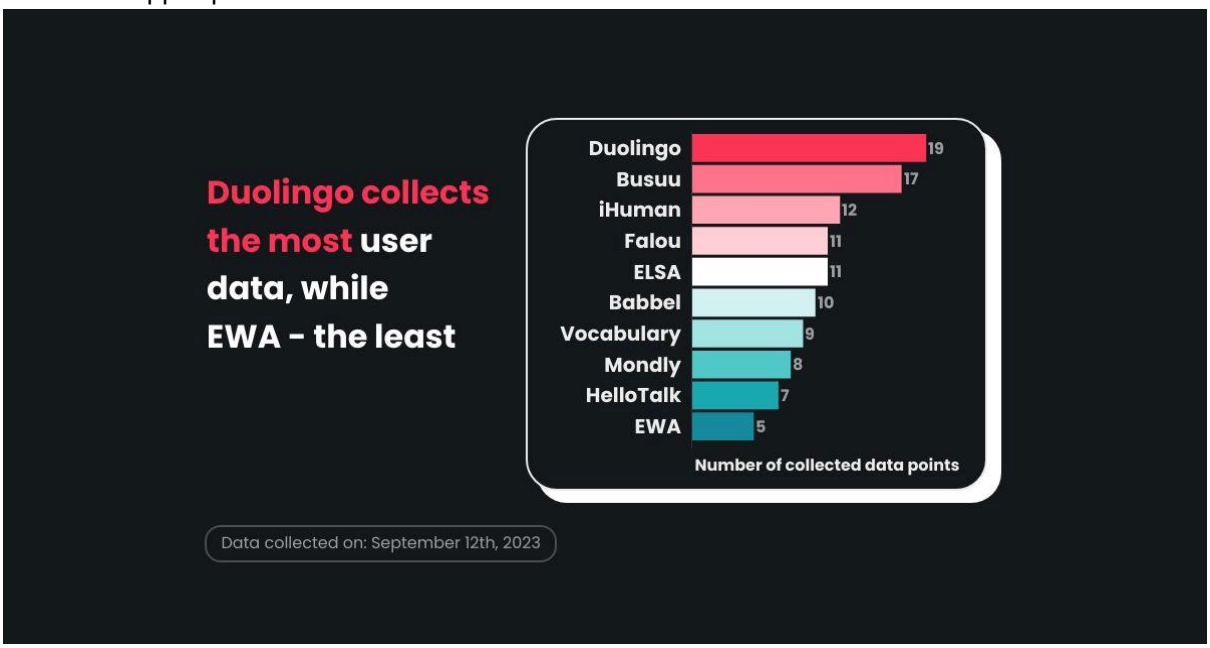

Fig 3.0

2. Data Ingestion:

- Tools: Apache Kafka, Google Cloud Pub/Sub.

- Process: Streams data in real-time from applications to data storage.

3.Data storage:

Duolingo uses Amazon S3 Cloud Storage for permanent storage, data is managed by Amazon DynamoDB, Amazon EMR in conjunction with Amazon's elastic block store (EBS) for short-term

storage, and Spark for often batch prediction calculations. Duolingo also uses Amazon Polly, a text-to-speech tool fueled by deep learning that integrates seamlessly into its applications, to give voice to a

-

- Duolingo utilizes the PyTorch deep machine learning technology on Amazon's Web Services (AWS).

4.Data analytics tools:

segmentation:

Duolingo relies on segmentation as a key data extraction strategy.

dashboard:

You may combine the graphs you start making for the data you are interested in into a personalized dashboard that displays several data points at once.

retention:

Duolingo offers a simple tool for displaying daily, weekly, and monthly retention in table and graph formats.

fig 3.1

funnel:

Custom funnel analyses can be created by combining numerous events in the user flow.

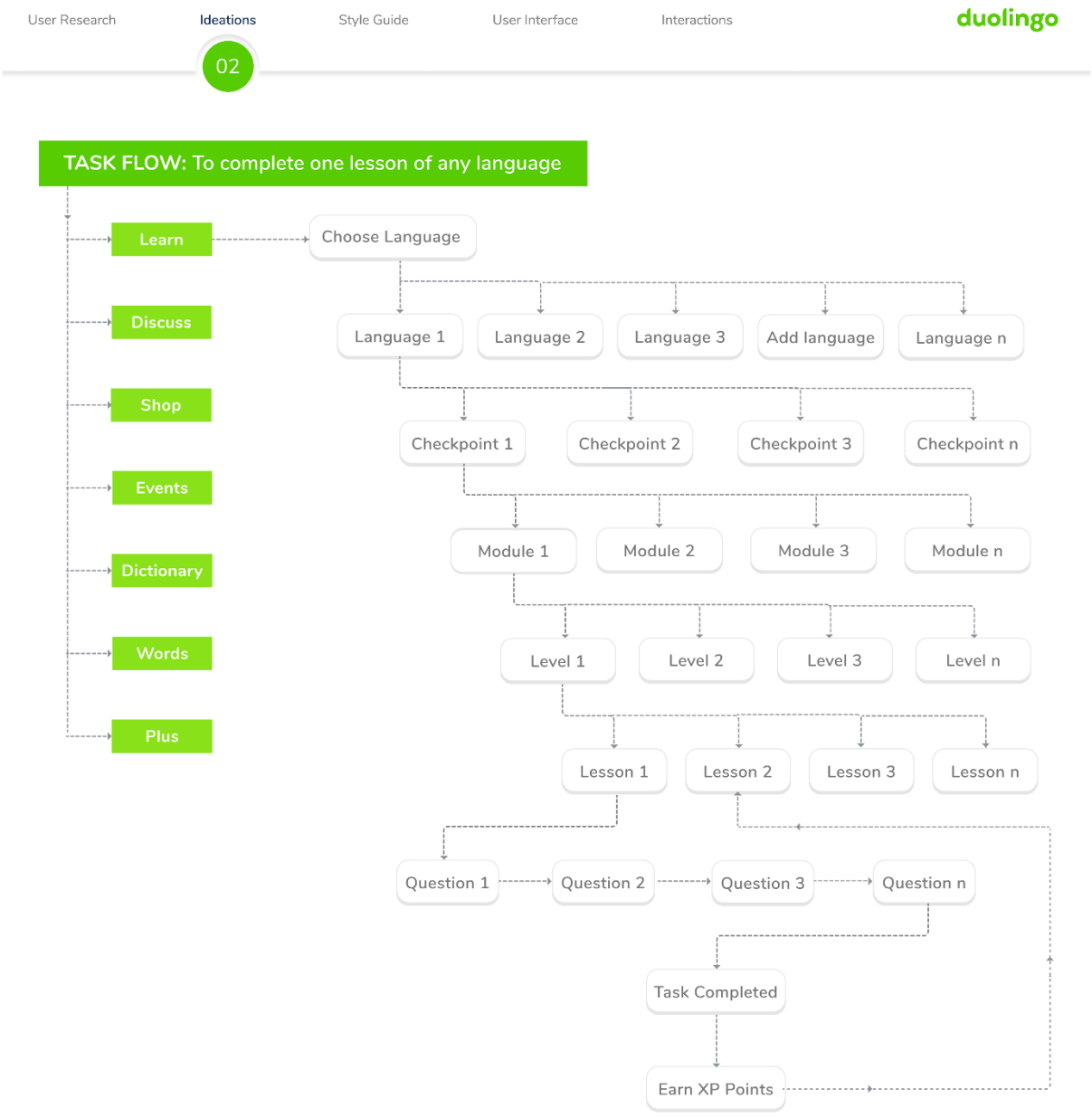

fig4.0

Tools and Services:

More advanced and specialized machine learning models were required for the kind of personalization Duolingo researchers were aiming for, as they discovered through A/B testing different strategies with users.
Once you have the data then we should start refining it with deep learning aspects.

1. A/B testing:

- purpose: 1. allows us to make data-driven product decision

2. If a change fails to work out in the way we had intended, it allows us to learn and improve.

2. Machine Learning:

Finding the remaining issues in a course might get more difficult as it matures and the major flaws are fixed. To address this issue, Duolingo's system method relied on a tried-and-true algorithm called logistic regression. A set of report features is fed into logistic regression, which then uses a scoring system to determine how much each feature is typical of high-quality reports. The total score is then used to determine the overall likelihood of a high-quality report.

3. Skill Tree Optimization:

- Technology: Graph Theory and Data Mining.
- Purpose: Structures lessons to create optimal learning paths.
- Data Used: User progression data and learning patterns.

Technologies Employed:-

1. BigQuery (Google Cloud Platform):

-A key element of Duolingo's data architecture is Google BigQuery, which allows the platform to effectively handle and analyse massive amounts of data. Duolingo can do quick SQL queries on large datasets with BigQuery, a serverless and highly scalable data warehouse. This is crucial for real-time data analytics and machine learning applications.

2. TensorFlow:

- Function: Machine Learning framework.
- TensorFlow is integral to Duolingo's strategy of providing a personalized, adaptive, and engaging language learning experience. By leveraging TensorFlow's powerful machine learning and NLP capabilities, Duolingo can continuously analyze user data and optimize content delivery, ensuring that each user receives the most relevant and effective learning experience possible.

3.Apache kafka

A key element of Duolingo's data architecture, Apache Kafka allows real-time data streaming throughout the site. Through real-time collection and processing of user activity data, Kafka allows Duolingo to rapidly observe interactions and gain insights. This real-time data flow is necessary to keep track of user behavior and system performance, which enables Duolingo to react fast to users' demands and dynamically optimize their learning experiences.

4.Tableau

Tableau plays a key role in Duolingo's data visualization strategy. It offers advanced tools for generating visual reports and dashboards that are used to evaluate user data and track performance metrics. These visual aids aid Duolingo in decision-making by making difficult facts understandable and useful.

Conclusion:

Handling and analysing large amounts of data is essential for providing personalized and creative user experiences in the quickly changing field of digital education. One example of this is Duolingo, which uses big data analytics to improve its offerings and hold onto its position as the market leader in language learning.

Duolingo's innovative use of big data analytics positions it as a leader in the digital education sector. By leveraging real-time user data, the platform delivers highly personalized and engaging learning experiences. Its continuous adaptation and refinement based on data insights not only enhance individual learning journeys but also set a high standard for the integration of big data in educational technology. As Duolingo grows, it remains dedicated to making language learning accessible, effective, and enjoyable for millions around the globe.

References

https://research.duolingo.com/

https://surfshark.com/research/chart/data-hungry-language-apps

https://aws.amazon.com/solutions/case-studies/duolingo-aws-is-how/

https://blog.duolingo.com/duolingos-secret-weapon-our-beautiful-and-powerful-analytics-tools/