Indian Institute of Management
Ahmedabad

IIMA/CISG0130

Data Modelling and Management for Big Data

CFX Inc, an e-commerce start-up based abroad, had built a large e-marketplace that allowed sellers and buyers to transact online. The firm had 30,000 sellers and aimed to increase the number to 50,000 in another year's time. In 2015, the company had around six million customers and anticipated their growth to triple in the next two years. In order to provide the best shopping experience to their growing customer base, the firm needed to collect, store and analyze different kinds of data (transactional, behavioural, syndicated and demographic) and improve customer shopping experience.

The company was in the process of identifying and designing suitable data management systems to sustain and manage its business growth. As part of this initiative, it hired a consultant to study its data management requirements, design a data model and offer implementation related recommendations.

The consultant interviewed key stakeholders in the organization and captured the following set of requirements:

User Registration Details

The system needed to maintain user (buyer and seller) registration details. The users primarily accessed and/or updated their information by providing their access credentials such as id and password. Some of the data captured by the system for buyers and sellers included:

Buyers: buyer identifier, name, age, gender, income, email id, phone, mobile, shipping address, billing address, preferred categories, credit card number and card type.

Sellers: seller identifier, company name, contact name, contact title, address, city, state, country, phone, email, payment method and rating.

Registration was mandatory for a seller, but, a buyer could purchase items from the site even without registration (i.e. as an anonymous user) just by providing payment, billing and shipping details at the time of check out. The system kept information about such buyers by creating a new identifier for them.

The buyers and sellers accessed the data purely based on their identifier for updating their profile. The seller data could be accessed by the buyers and the CFX employees as well. Their needs were primarily to access seller information by their identifier, name, city, rating or payment method.

In addition to the buyers and sellers, who were the primary users, the system also kept track of CFX's marketing affiliates and partner related details.

Prepared by Professor Srikumar Krishnamoorthy, Indian Institute of Management, Ahmedabad.

This case is based on the analysis of an online retail business in India from secondary data sources.

Cases of the Indian Institute of Management, Ahmedabad, are prepared as a basis for classroom discussion. They are not designed to present illustrations of either correct or incorrect handling of administrative problems.

'Look-to-Click' Phase

The company received around a million visitors per day and on an average a customer viewed eight pages per visit. The average time spent by a visitor in the site was around nine minutes. A sizeable volume of click data was generated by the system due to large number of customer visits. With the increasing growth in e-commerce business and the popularity of CFX Inc, these numbers were likely to go up significantly in the next couple of years.

The company intended to increase the total number of sellers to provide a large variety of offering to its customers, while monitoring the quality of merchandise sold by the sellers. In addition, the company planned to strengthen its marketing affiliate program to drive more traffic to the online store and improve overall profitability. At that time the company generated roughly 12% of its sales through its affiliate program.

The management expected to build a data infrastructure that supported a peak load of 30,000 reads per second. Such an infrastructure would allow the users to have a richer shopping experience. At the same time, an analyst within the company performed real-time analytics on customer navigation patterns and offered personalized user experience thereby increasing conversion rates.

The click data needed to be persisted in permanent storage for performing advanced offline data analytics to identify interesting patterns or trends in the data. Some of the advanced analysis requirements included customer segmentation, campaign effectiveness analysis, market basket analysis, and identification of next best offer for a given customer.

The click data generated by the system in the form of weblogs included session information, date and time of click, nature of click (product page view, campaign or offer selection, and search request), URL details, location information, device or app details, referrer details, request status and optionally user information. Some of the key analysis requirements identified from the interviews with stakeholders were: (a) number of pages visited by state, city and user in a given time period, (b) total number of hits by a given referrer/affiliate, (c) total number of page views for a given user by time period, (d) number of page views by device type or operating system, (e) number of page views or amount of time taken by the user from click to basket / look to purchase and (f) number of users who added items to the shopping cart but did not buy them in the same session.

'Click-to-Basket' Phase

The users of the site added or removed items from the shopping cart as they navigated across product pages and were allowed to create a wish list of items for future purchase. The user generally accessed the wish list or cart to review orders and make suitable updates. It was possible that the information in the cart or wish list was stale. For instance, a seller might update the pricing or availability details after the user had placed the items on the cart. Limited amount of incorrect information like minor changes in price or product catalogue details was acceptable to the user. However, the user generally was dissatisfied if the item was initially shown as available and the order got executed, even though the item got sold to another user in the intervening period. This was likely to happen if there was a delay between customer placing items in the cart and the actual checkout. Therefore, there was a need to update the cart information during the checkout phase or at least notify the user to maintain appropriate customer satisfaction levels.

Each user typically had just one shopping cart or wish list. The shopping cart was abandoned automatically by the system beyond a pre-configured time period (say, after 3 days).

'Basket-to-Purchase' Phase

The cart abandonment rate was estimated to be 60% to 80% for online retailers which meant that only a fraction of the shopping carts got converted to actual orders and transactions. Even a small improvement in abandonment rate could lead to significant increase in profitability for the online retailer or e-marketplace.

On an average the company received 25,000 orders per day with a peak load of 40,000. Each order could have one or more transactions. The system needed to maintain consistency of transactions with minimal latency during the order placement and payment process.

The order fulfilment process involved updates of the order by different stakeholders inside and outside the organization. The system was expected to maintain consistency of updates in the order fulfilment pipeline. A customer was likely to request for order status information on a regular basis. Therefore, providing timely and correct order status information was critical to ensure customer satisfaction.

Business users in CFX Inc. were interested in understanding: (a) total size of a given customer order, (b) average value of the order, (c) number of orders by shipping state and city, (d) list of top-10 products by value or volume, (e) most common payment methods and (f) total number of orders placed by customers per month.

Customer Recommendations

The company needed to track customer product purchase patterns to offer personalized recommendations (e.g. users who purchased a book on 'Data management' also purchased books on 'Databases', 'Data analytics' and so on). It also planned to conduct similar analysis with the help of customer click navigation information. The prevailing e-marketplace did not provide such recommendations to the customer. There was a need for the system to capture such product purchase relationship details and allow performing advanced analytics.

User Engagements

The e-commerce platform provided facilities for a customer to engage actively with the community of buyers. A customer could offer reviews of products he had purchased or rate them; write comments on reviews by other users; express likes and vote for customer reviews. This data was highly unstructured and completely user-generated. Such consumer-generated content was often used by potential buyers for making suitable purchase decisions. Therefore, it was important for the e-marketplace to organize the content in a manner suitable for potential buyers to make better product choices. This entailed the need for conducting advanced analytics to determine better organization of user-generated content.

The number of data accessed (or read) was likely to be significantly large given the large number of visitors and the amount of time a user spent on the site. As per estimates, the site was likely to receive 3,000 to 7,000 accesses per second. The system had to be designed to handle a peak load of 15,000 accesses per second.

Following analytical queries were of interest to the management: (a) top-rated products/sellers in the last one month (b) list of product reviews by recency or helpful votes and (c) total number of reviews/votes/comments for a given product.

Seller Inventory and Pricing Management

CFX Inc's product catalogue had five million products across 600 product categories and 30,000 sellers. The system maintained basic product details such as SKU, name, description, category, available quantity and reorder quantity.

The system needed to maintain information about different categories of products. Each category of product might have different set of attributes. For example, 'book category' had attributes like author, title, publisher, publication date and release version. On the other hand, a 'music category' had attributes like artist, title, album name and track names. It was also likely that the list of attributes, their types and names varied from one seller to another. As more and more new categories of products and sellers were added on a regular basis, it was not possible to pre-determine or model the attributes in advance.

The inventory and price of items were updated on a regular basis by the sellers. The system had to handle large volumes of updates made by the seller in real-time without any disruption to the customer experience. The system was expected to handle 500 average updates per second (and handle a peak load of 1,000 updates during festive seasons) with a 95% latency of 10 milliseconds (ms).

Data Management Issues

CFX Inc used a relational database for maintaining detailed information about its customers, sellers, products, click and purchase histories. Over time, the underlying relational data model used by the company had grown fairly complex. With exponential increase in business volumes, the company was facing severe challenges in scaling their database infrastructure. Besides, there was an increasing need for performing real-time analytics to improve customer online experience through personalization.

The company's data management team constantly monitored their data access/query performance to provide superior experience to their customers. The team had observed a significant degradation in query latencies from an average of 8-10ms for pricing updates to as high as 40ms and in some cases even 100ms. Similar performance degradation was also observed during order processing and billing. Prior empirical studies demonstrated that significant increase in latencies often led to customer permanently leaving the e-marketplace.

Future of Data Management

With increasing business volumes, greater customer expectations, and the consequent need for performing advanced batch or real-time analytics, the management was forced to take immediate action and explore infrastructure scaling alternatives. The management understood the need to keep in mind the scalability, flexibility, cost-effectiveness and their applications consistency requirements while designing the new solution.

Vertical scaling of their prevailing database infrastructure and/or database storage clustering were the immediate options that were readily available. However, the management was seriously considering newer big data management solutions. Given their lack of expertise in the field, the management sought expert guidance of the consultant.

The consultant was to guide them on the following aspects:

a. The data management alternatives that are suitable for CFX Inc
b. The trade-offs involved in exploring these alternative solutions
c. Typical costs, benefits and ROI of these alternative solutions
d. How should CFX Inc handle the modelling data in the new environment?
e. What were some of the past success stories and the key learnings for CFX while adopting these new data management solutions?
f. What were the implementation considerations?