Analyzing Steam User's Behavior Using Data-Warehouse Method

Article *in* Human Behavior and Emerging Technologies · February 2025

3 authors, including:

Aldi Patria
Binus University
1 PUBLICATION   **0** CITATIONS

Analyzing Steam User's Behavior Using Data-Warehouse Method

Yusuf Fahry, Aldi Patria, and Jaka Prana Atmasudirja

Abstract—We present a guide on how to analyze Steam gaming network using a data-warehouse method with open-source tools. With hundreds of thousands of data comprising the app's information and the historical data of the app's active session, we want to examine user behavior across the dimensions of game's developers and publishers, genres affinity, and money expenditure. Gamer behavior is highly tied with a wide diversity and characteristics by distributions. Thus, finding out how to build the right data warehouse will give a significant impact on almost everything such as query performance, ease of analysis, and maintenance.

Index Terms—business intelligence, big data, data warehouse, costumer analysis

I. INTRODUCTION

PC gaming has existed since the dawn of desktop computers, but rather than fade into oblivion as mobile devices and game systems took over, the industry has remained a juggernaut.

By the end of 2016, PC gaming sales will reach $29 billion worldwide, compared to $28 billion for console gaming platforms like Xbox and Play Station [1]. That projection is thanks, in part, to the growing legions of PC gamer in developing countries like India and China. PCs are also the chosen medium for hosting innovative virtual reality technologies like Oculus Rift.

Few companies have been able to benefit from this trend as well as Valve Corporation and its digital subscription platform, Steam. And while consumer choices would seem to favor the company, it helps that Steam had a first-mover advantage. Tack on the company's willingness to listen to its customers and adapt along the way and Valve would seem to be a business case study with which any entrepreneur should become familiar.

In 2011, it is estimated that Valve's value was somewhere between $2 billion to $4 billion, with Steam generating roughly $400 million every year [2]. Recently, Steam hit a record-breaking 10 million concurrent users. The platform also now has more than 100 million users. [3]

Steam's success teaches the business to identify and build a relationship with the customers. Understanding the consumers brings about growth and development. The business success requires the ability to know the weakness and strength [4]. An organization is advised to work on the strength instead of fixing the weakness. They should be able to adapt to all condition in order words tackle all problems as they come.

Yusuf Fahry, Aldi Patria, Jaka Prana Atmasudirja are with School of Computer Science, Bina Nusantara Univesity, e-mail: yusuf.fahry@binus.ac.id, aldi.patria@binus.ac.id, jaka.atmasudirja@binus.ac.id

Steam provides a unique opportunity for analysis. Because Steam offers an open API for their platform, we are able to collect an exhaustive measurement of all users and games. Steam provides data on friendships, game ownership, and playtime data, but not access to more detailed game statistics. We thus necessarily focus on a general characterization of aggregate behaviors among Steam users.

So what are gamer really like? Using this dataset, we investigate questions concerning gamer behavior, such as how many gamer play, what types of games they play, and how are discounted items affecting the game they play.

II. THEORY

A. Big Data and Data Warehousing

The importance of data-warehouse in an organization tends to explain why data warehouse needed in an organization. Since a data warehouse reflects the business model of an enterprise that makes is an important aspect of an organization. The following are some importance of data-warehouse:

1) Repository for historical information for comparative and competitive analysis
2) Ability to enhanced data quality and completeness.
3) Real-time consolidation of financial data becomes practical.
4) The IT costs and staff dedicated to reporting are greatly reduced.
5) Allow business process redesign and align with business strategy.
6) Give end users freedom to carry out wide-ranging analysis in various manners.
7) Simplify the process of data access.
8) Identify market trends.
9) Reduce operation costs.
10) Allow business process redesign and align with business strategy.

Out of all number 10 is the most important one for businesses.

A data warehouse system (DWH) is a kind of decision support system (DSS) that is data-driven. It can support in narrowing an organization's decision-making to be more accurate. For example, a repository of sales data can help a company building their targeted marketing model in order to reduce their marketing cost because of mistargeted marketing [5], finding the best plan for improvements based on previously captured historical data of sales performance, and so many more.

When it comes to designing a data warehouse, there are two famously known guidances that can be followed. We can divide it into transactional (OLTP) and analytical (OLAP). In

general, we can assume that OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it [6].

1) **OLTP** (*On-line Transaction Processing*) is characterized by a large number of short online transactions (INSERT, UPDATE, DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by the number of transactions per second. In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

2) **OLAP** (*On-line Analytical Processing*) is characterized by a relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In the OLAP database, there is aggregated, historical data, stored in multi-dimensional schema (usually star schema).

TABLE I
SUMMARY OF THE MAJOR DIFFERENCES BETWEEN OLTP AND OLAP
SYSTEM DESIGN [6].

Feature	OLTP System	OLAP System
Source of data	Operational data	Consolidation data
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data are	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Standardized and simple queries. Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Very fast	Depends on the amount of data involved
Space Requirement	Small	Larger due to the existence of aggregation structures and history data
Database Design	Highly normalized	Typically de-normalized with fewer tables
Backup and Recovery	Backup religiously	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

B. Customer Analysis for Strategic Planning

Plans are needed to advance relationships with chosen customers and potential customers, and preparation needs to precede planning [7]. Some have suggested that companies that are able to identify their customers, differentiate them by need and value, and have the potential to interact with each customer uniquely as an individual, while changing aspects of their products or services to serve individual customers, would provide some evidence of preparation.

Relationship management by its very nature requires more than a near-term orientation and comprises a vision of the organization as it could be. As such, relationship management will require more than passive support from the CEO and board to enable the strategy to be pursued; it will require active involvement.

Some features a company has may be valuable because they attract new customers, while others are valuable because they retain existing customers. For example, having a strong user community can be a powerful feature for retaining gamer. On the other hand, if the community behaves aggressively toward rookie players, a strong community may scare new users away.

An added challenge is that consumers themselves may evaluate features differently when they are making an initial choice between alternatives versus when they are evaluating their satisfaction with a consumption experience. Typically, adding more features or amenities increases the likelihood that consumers will choose a company's goods and services [8], each feature added gives the customer one more reason to choose the product. For example, when customers choose a hotel, airline, or theme park, their choices may be influenced by amenities or star attractions in addition to core attributes such as location or price. To maximize initial choice, managers often decide to offer more amenities [9].

III. METHODOLOGY

A. Data Collection

Valve provides a REST API called the Steam Web API we can use to retrieve information about user's profiles, game informations, playtimes, and more. Since we are going to analyze the gamers playing behavior, these are the API we mainly use:

steam.steam.game
- API: ISteamApps - GetAppList - v2
- Descriptions: Gets the complete list of public apps
- Schedule: Daily
- Output: *steam.steam.game.json*

steam.steam.gameCurrent
- API: ISteamUserStats - GetNumberOfCurrentPlayers - v1
- Descriptions: Gets the total number of players currently active in the specified app on Steam
- Schedule: Hourly
- Output: *steam.steam.gameCurrent.json*

While collecting the data we tried not to break the terms of conditions in using the public API. Requesting a lot of data in a single time will leave a heavy job for Steam's infrastructure. Thus, we split the job. Say there are 20 thousand games available on Steam, we split the job of calling the API to get the details of each game. For example, if we split it by 10 jobs, that means each job will retrieve two thousand of JSON data. For each job also we give a break of two minutes for every 100 data retrieved.

Every data will be then processed with ETL for cleansing, extraction, and many more. The output will be stored as CSV

file and Parquet file. Where CSV format will be used in COPY FROM method insertion to database, and Parquet file is to provide a small size of data that can be processed in case we want to do a full analysis of the whole data. For the database itself we are using PostgreSQL because of its variety in the supported data type.

B. Data Visualization and Analysis

Data visualization refers to the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization is an accessible way to see and understand trends, outliers, and patterns in data. Our eyes are drawn to colors and patterns [10]. We can quickly identify red from blue, square from circle. It is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers.

In order to visualize data correctly, these are the advised step can be followed [11] :

1) **Understand the context**, the first thing to do when faced with a visualization challenge is to make sure we have a robust understanding of the context and what we need to communicate.
2) **Choose an appropriate display**, once we've identified the data we want to show, next comes the challenge of determining how to best visualize it.
3) **Eliminate clutter**.
4) **Draw attention where you want your audience to focus**.
5) **Tells a story**, make the visualization telling a story to the reader.

The data we are going to use is a continuous or historical data. In order to visualize correctly, we need to have two variables that will act as the measurement and the dimension. In this case, we decided to choose variables that represent date and numbers.

C. Database Design

We are following the style in star schema database Figure 1. Along with the table description Tables II, III and IV .

TABLE II
TABLE DESCRIPTION FOR DIM_TIME

Column	Description
DATA_DATE	Timestamp indicating the data's time
HOUR	Hour part of DATA_DATE
DAY_OF_MONTH	Day of month
WEEK_OF_MONTH	The week number in a month period of DATA_DATE
DAY_OF_WEEK	The day number in a week period of DATA_DATE
YEAR	Year part of DATA_DATE
MONTH	Month part of DATA_DATE

IV. RESULTS

From the daily active session, we see that there was a peak on 16th July 2018 Figure 2 . Although the date was near the
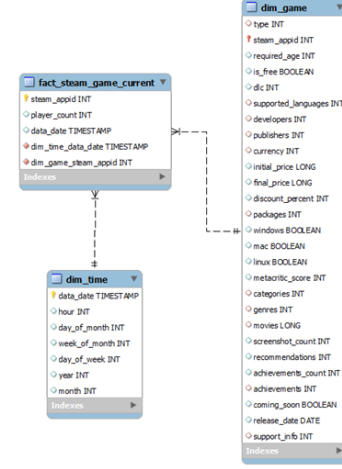

Fig. 1. Star schema design for the database

TABLE III
TABLE DESCRIPTION FOR FACT_STEAM_GAME_CURRENT

Column	Description
STEAM_APPID	ID that represents the app
PLAYER_COUNT	Numbers of the active session (player)
DATA_DATE	Timestamp

TABLE IV
TABLE DESCRIPTION FOR DIM_GAME

Column	Description
TYPE	Type of app (software, game, etc.)
STEAM_APPID	ID that represents the app
REQUIRED_AGE	App's minimum age
IS_FREE	Is a free item or not
DLC	Numbers of downloadable content the app has
SUP-PORTED_LANGUAGE	In-game supported language
DEVELOPERS	The developer of the app
PUBLISHERS	Publisher of the app
CURRENCY	App's selling currency
INITIAL_PRICE	Price of the app before discount
FINAL_PRICE	Price of app after discount
DIS-COUNT_PERCENT	Discount percentage
PACKAGES	List of packages/bundle where the app is included
WINDOWS	Whether the app can run on Windows platform or not
MAC	Whether the app can run in Mac platform or not
LINUX	Whether the app can run on Linux platform
META-CRITIC_SCORE	Professional reviewer score towards the app
CATEGORIES	Category of the app
GENRES	App's list of genre
MOVIES	Count of movies the app is having in the store page
SCREEN-SHOT_COUNT	Count of screenshots the app is having in the store page
RECOMMENDA-TIONS	Number of players recommending the app
ACHIEVE-MENTS_COUNT	Number of achievements (feature) the app is having
COMING_SOON	Whether the app is published or not
RELEASE_DATE	Date of publish

US independence day moment, the graph shows us that the most active game was not discounted Figure 3 or free apps Figure 4. Still, the most popular game being played is the non-discounted game. This might indicates that money is not really a problem for gamer that resides in the Steam community.
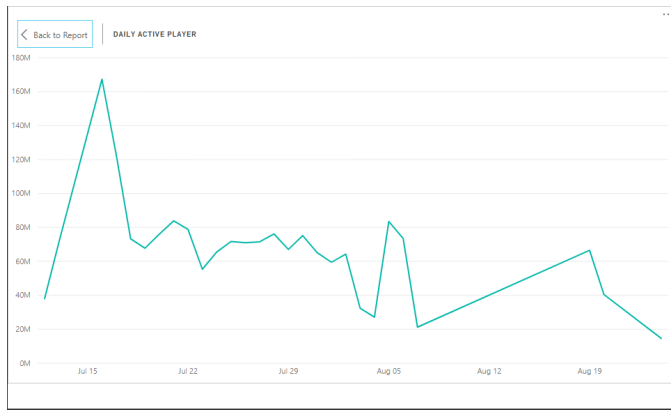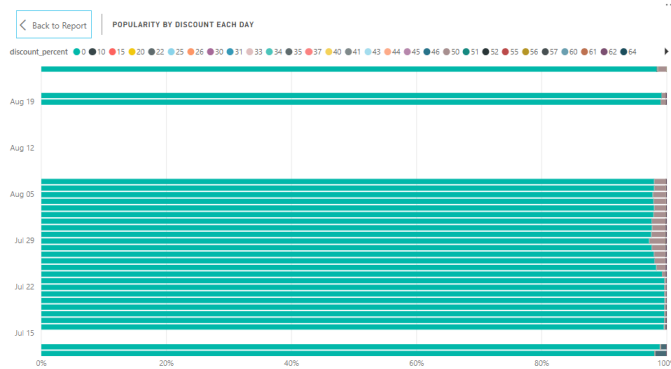

Fig. 2. daily active player
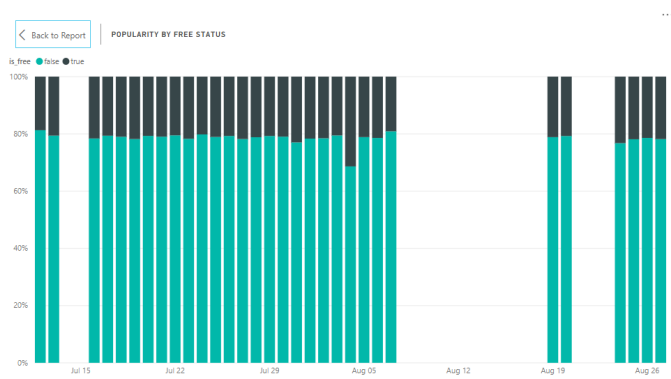

Fig. 3. popularity by discount

Fig. 4. popularity by free status

Digging deeper into it, Figure 5 we see that mostly it is dominated by developer Valve Corporation in the first place, followed by PUBG Corporation in the second place. Furthermore, if we break it down, even more, we'll see that the daily top 3 are dominated by PUBG's Playerunknown Battleground and Valve's Dota 2 and Counter-Strike. Figure 6
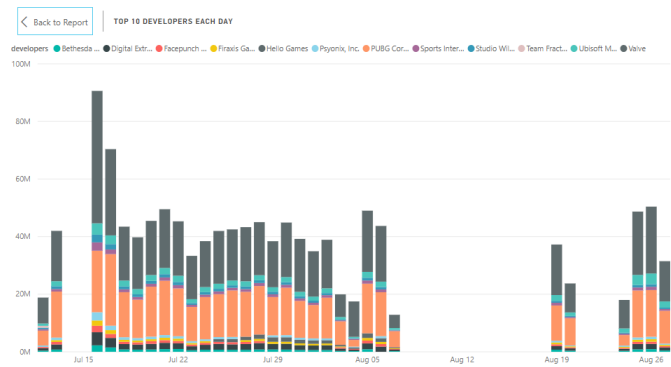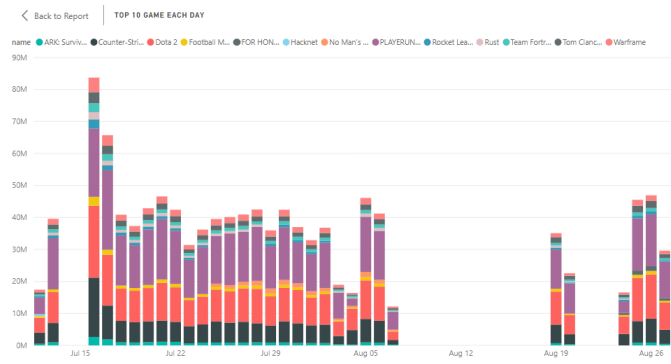

Fig. 5. top 10 developers each day

Fig. 6. top 10 games each day

Playerunknown's Battlegrounds is a last-man-standing multiplayer arena shooter. It blends elements of traditional survival games like DayZ with the fast-paced action and leaderboard of something like Counter-Strike. Around 100 players parachute onto the eight-square kilometer map, scavenge for supplies and fight to the death. The last one alive gets a "chicken dinner" filled with in-game currency. That is to say they get the pride of being #1 and a few virtual coins.

This isn't a new game type. The battle royale genre has been around since roughly 2013.

Dota 2 is considered as an action real-time strategy (ARTS) game or also known as a multiplayer online battle arena (MOBA), in which a player controls a single character in a team who compete versus another team of players. The objective is to destroy the opposing team's main structure with the assistance of periodically-spawned computer-controlled units that march forward along set paths.

Playerunknown Battleground and Dota 2 are sharing a gene between them. Both are a competitive game, pushing the player to defeat the opposing team or player. It is a place where people are trying to win over others.

Previous research found that people like competition. Both in men and women, the desire to win over other people is significantly high. People are always trying to make improvements of themselves. 71% of 151 participants aged 18-35 stated that they find satisfaction in exceeding their previous performances even they are not outperforming their others [12]. It is also stated that their motivation to put forth effort is greater in a competitive situation.

V. CONCLUSION

A data warehouse can be a role player in driving someone's success. With the right preparation, planning, and implementations, it can lead to meaningful insights for the company. In this case, we applied an OLAP design to analyze massive data to gain insight about the customer. We found that Steam users mostly are people that tend to like a competition against other people, wanted to sense the beauty of winning over others.

This simple insight can help many people. Most importantly for game developers, this can help them in planning the style of their next game.

REFERENCES

[1] N. Statt, "Playing games on the PC is making a comeback," 2015. [Online]. Available: https://www.cnet.com/uk/news/playing-games-on-the-pc-is-making-a-comeback/

[2] O. Chiang, "Valve And Steam Worth Billions," 2011.

[3] E. Makuch, "Steam Reaches 100 Million Users And 3,700 Games," 2014.

[4] W. C. Buboltz, M. Miller, and D. J. Williams, "Content analysis of research in the Journal of Counseling Psychology (1973–1998)," *Journal of Counseling Psychology*, vol. 46, pp. 496–503, 1999.

[5] V. Marotta, K. Zhang, and A. Acquisti, *Who Benefits from Targeted Advertising?*, 2015.

[6] I. E. S. Limited, *Introduction to Database Systems, in: Introduction to Database Systems*. India, India: Pearson, 2008.

[7] A. Ries, J. Trout, and P. Kotler, "Positioning: The Battle for Your Mind," in *Positioning: The Battle for Your Mind*. McGraw-Hill Education, January 2001.

[8] R. T. Rust, D. V. Thompson, and R. Hamilton, "https://hbr.org/2006/02/defeating-feature-fatigue," 2006. [Online]. Available: https://hbr.org/2006/02/defeating-feature-fatigue

[9] R. D. van Oest, H. J. van Heerde, and M. G. Dekimpe, "RETURN ON ROLLER COASTERS: A MODEL TO GUIDE INVESTMENTS IN THEME PARK ATTRACTIONS ," *Marketing Science*, vol. 29, no. 4, pp. 767–780, 2010. [Online]. Available: 10.1287/mksc.1090.0553

[10] R. Mehta and R. J. Zhu, "Blue or red? Exploring the effect of color on cognitive task performances." [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19197022

[11] C. N. Knaflic, "Storytelling with Data: A Data Visualization Guide for Business Professionals," in *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons, November 2015.

[12] R. E. Franken and D. J. Brown, "Why Do People Like Competition? The Motivation for Winning, Putting Forth Effort, Improving One's Performance, Performing Well, Being Instrumental, and Expressing Forceful/Aggressive Behavior," *Personality and Individual Differences*, vol. 19, no. 2, pp. 175–184, 1995. [Online]. Available: 10.1016/0191-8869(95)00035-5