Lappeenranta University of Technology
School of Business and Management
Degree Program in Computer Science

Bachelor’s thesis

Kalle Kareinen

BIG DATA IN VIDEO GAMES

Examiner(s): M.Sc. (Tech) Antti Herala

Supervisor(s): M.Sc. (Tech) Antti Herala

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto
School of Business and Management
Tietotekniikan koulutusohjelma

Kalle Kareinen

Big Data videopeleissä

Kandidaatintyö

2017

27 sivua, 2 kuvaa, 5 taulukkoa

Työn tarkastajat: DI Antti Herala

Hakusanat: big data, videopelit, videopeliteollisuus
Keywords: big data, video games, video game industry

Big Datasta on tullut oleellinen käsite Internet-yhteyksien levitessä ja kehittyessä. Tässä työssä tutkittiin Big Datan käyttöä videopeleissä. Työn tavoitteena oli selvittää, millaisia käyttötarkoituksia Big Datalle on videopeleissä ja kuinka pelikehittäjät hyötyvät Big Datasta. Työssä kerrotaan yleisesti Big Datasta ja siihen liittyvistä teknologioista. Työssä käydään läpi peliteollisuuden kehittyminen nykyiseen tilaansa ja millaisia ongelmia nykypeleillä on. Peliteollisuudesta käsitellään erilaisia rahoitusmalleja. Big Datan hyödyntämistä videopeleissä käsitellään erilaisilla esimerkeillä yrityksistä Blizzard, Supercell ja Zynga. Pelien ongelmat liittyvät erityisesti pelisuunnitteluun ja erilaisten ominaisuuksien määrään. Big Data auttaa kehittäjiä balansoimaan, testaamaan ja monetisoimaan heidän pelejään. Big Datalle löytyy tulevaisuudessa todennäköisesti vielä enemmän käyttötarkoituksia videopeleihin ja niiden kehitykseen liittyen.

ABSTRACT

Lappeenranta University of Technology
School of Business and Management
Degree Program in Computer Science

Kalle Kareinen

Big Data in video games

Bachelor’s Thesis

2017

27 pages, 2 figures, 5 tables

Examiner: M.Sc. (Tech) Antti Herala
Keywords: big data, video games, video game industry

Big Data has become a relevant concept as Internet connections have spread and developed. This thesis researched the usage of Big Data in video games. The goal for this thesis was to find out what are the use cases of Big Data in video games and how the game developers benefit from Big Data. This thesis presents Big Data and technologies related to it in general. This thesis goes through the development of video game industry to the current state of it and what kind of problems do the games have today. The monetization models of video games are presented as one part of this thesis. Usage of Big Data in video games is discussed with examples of companies Blizzard, Supercell and Zynga. The problems of developing video games relate mostly to game design and the number of different features. Big Data helps developers to balance, test and monetize their games. In the future, Big Data will most likely have even more use cases in video games and their development.

INDEX

1 INTRODUCTION 3

1.1 OVERVIEW 3

1.2 GOALS AND DELIMITATIONS 3

1.3 STRUCTURE OF THE THESIS 4

2 BIG DATA 5

2.1 DEFINITION 5

2.2 DATABASES 6

2.2.1 Relational databases 6

2.2.2 Non-relational databases 7

2.3 FRAMEWORKS 8

2.3.1 Apache Hadoop 9

2.3.2 High-Performance Cluster Computing 10

3 EVOLUTION OF VIDEO GAMES 12

3.1 EVOLUTION OF GAME TYPES 12

3.2 EVOLUTION OF DEVELOPING VIDEO GAMES 13

3.3 EVOLUTION OF MONETIZATION MODELS 13

3.4 MOST COMMON PROBLEMS OF GAMING INDUSTRY 15

4 USAGE OF BIG DATA IN VIDEO GAMES 18

4.1 BALANCING 18

4.1.1 Player character balancing 18

4.1.2 Game environment balancing 19

4.2 MONETIZATION 21

4.3 TESTING 21

5 DISCUSSION 24

6 CONCLUSIONS 25

REFERENCE 26

LIST OF SYMBOLS AND ABBREVIATIONS

CMS	Content Management System
ECL	Enterprise Control Language
F2P	Free to play
HFDS	Hadoop Distributed File System
HPCC	High-Performance Cluster Computing
IAP	In-app purchase
ISAM	Index Sequential Access Method
JSON	JavaScript Object Notation
LAN	Local Area Network
MOBA	Multiplayer online battle arena
NoSQL	Not only SQL
P2P	Pay-to-play
PTR	Public Test Realm
RDBMS	Relational Database Management Systems
SQL	Structured Query Language

1 INTRODUCTION

1.1 Overview

Gaming industry has changed a lot during the last 20 years. New game types have emerged using the possibilities offered by fast and reliable connections spreading around the world. Nowadays gaming consoles and computers are usually connected to Internet during the gaming sessions even if the connection is not necessarily needed for playing the game.

Thanks to the developments in both Internet connections and hardware, it is easier for gaming companies to collect useful game data. It is also possible for gaming platforms such as Origin to create player profiles from their users that can be used for advertising new games and products directly for the user [1]. Big data helps companies to use all the incoming, excessive amount of data to make better decisions in both game development and business side of things [1].

Big Data offers gaming companies an option to make decisions based on fresh, almost real-time data collected automatically from their actual customers. When a new patch hits the servers and people start playing it, the results can be carefully followed using big data and adjustments can be made on-the-fly if the changes seem too dramatic for one direction or another. With the Big Data, it is easier to isolate the cause of the problem and therefore it is easier to focus on the correct things in game development. [2]

1.2 Goals and delimitations

This Bachelor's thesis uses literature and news articles to define the following topics:

- How is big data used in the video game industry?
- What are the benefits of using Big Data in the video game industry?

There will not be any surveys to gaming companies about their usage of Big Data because there are not enough resources for doing a thorough survey. The problem with survey would also be that companies could be reluctant to answer how they use all their data.

1.3 Structure of the thesis

This Bachelor's thesis consists of three main sections in addition to the Introduction section in the beginning of the thesis and both Discussion and Conclusions in the end of the thesis.

Section 2 introduces Big Data, what it is and how it can be used. Section 2.1 explains the definition of Big Data. Section 2.2 introduces databases that are related to the Big Data. Section 2.3 introduces the most-known Big Data software.

Section 3 introduces evolution of video games from offline single player games with a single launch and no updates to online multiplayer games of today. Section 3.1 introduces evolution of different game types. Section 3.2 goes briefly through the evolution of game development. Section 3.3 introduces how the Internet age has changed monetization models of games. Section 3.4 defines the common problems of developing modern multiplayer games.

Section 4 introduces how Big Data is used in the continuing development process of multiplayer games. Section 4.1 focuses on balancing a multiplayer game using Big Data. Section 4.2 introduces handling the monetization using data-centric approach to the game development. Section 4.3 lists examples of how testing can be done using Big Data.

Section 5 includes discussion about the study and the subjects. Section 6 introduces the conclusions that have been made from this study.

2 BIG DATA

2.1 Definition

Big Data does not mean any database that just happens to be a massive set of data. One of the most common definitions of Big Data consists of so-called “3Vs” model [3]. The model was introduced 2001 by Doug Laney and it consists of the following [3]:

- Volume
- Variety
- Velocity

However, during the 21st century it has been discussed whether “3Vs” is enough for the definition. Other “Vs” have been added to the definition by different authors, for example Veracity and Value [4]. To keep this thesis simple enough, 3Vs model with Volume, Variety and Velocity will be used as the definition of Big Data while acknowledging that other Vs exist and could be added to the definition.

Volume in the definition of Big Data means that there is simply too much data to store and handle without considering and handling it as Big Data. For example, according to Michael Frampton’s book Big Data Made Easy (2015), eBay has 40 petabytes (40 million gigabytes), of semi-structured and relational data stored in its system. [3]

Variety in the definition of Big Data means that data is in different formats. There could be any kinds of files, for example text files, images or videos to mention the most common ones [3]. Social media services like Facebook, Twitter and Instagram are good examples of services that have a huge variety of data coming in all the time.

Velocity in the definition of Big Data means that massive amount of data is flowing in continuously and it must be identified and categorized quickly [3]. When velocity is big enough, companies can get reliable real-time information and make decisions based on it [3]. If data is not handled fast enough, the commercial value of it is not maximized.

2.2 Databases

According to Guy Harrison, there has been three database revolutions by far [5]:

1. 1950-1972: Pre-Relational databases
2. 1972-2005: Relational
3. 2005-: The Next Generation

Pre-Relational databases consist of tools that are irrelevant today, such as Magnetic Tape and Magnetic Disk. Between these physical object databases and relational database management systems (RDBMS), there were Index Sequential Access Method (ISAM) by IBM which was electronic database but it could be accessed only by the one running application and therefore it is not counted as a database management system (DBMS). [5]

When database types are considered for Big Data, it is enough to talk about two of the latest revolutions since both have created databases that are still up and running [6].

2.2.1 Relational databases

Relational databases have been used for storing data even before Internet existed and they are still doing well in terms of popularity. For example, the most popular Content Management System (CMS) WordPress uses MySQL [7,8]. Relational databases such as MySQL have a clear and strict structure and it is defined accurately what kind of data can be saved, how much and in which column. Database operations, such as creating new tables, inserting new data and modifying it, are done using Structured Query Language (SQL) [9].

Table 1. Example of MySQL table, adapted from [10]

Id (Integer)	**Name (Varchar (10))**	**Age (Integer)**	**Occupation (Varchar (20))**
1	Adam	37	Lumberjack
2	Bernie	42	NULL
3	Charlie	50	Taxi-driver

Relational databases can store enormous volumes of data and therefore they still are very useful for content management systems like WordPress. However, performance of RDBMS is not good enough for the velocity of Big Data and RDBMS has too strict data schemas for the variety of Big Data [11]. When it comes to scalability, RDBMS is not very cost-effective compared to Big Data databases and therefore volume of Big Data could be possible to keep up but it could be too expensive [11].

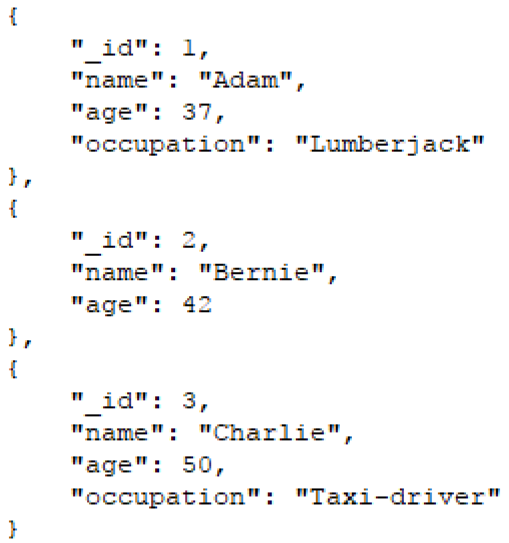
2.2.2 Non-relational databases

Non-relational database management systems, also known as Not only SQL (NoSQL) are the The Next Generation of databases [5]. They are used for Big Data because they can receive and handle massive amounts of data better when all the three 3Vs, volume, velocity and variety of the data are huge [12]. NoSQL databases do not have as strict and predefined database structure as relational databases [12]. There are four types of different NoSQL data store types [5].

Table 2. NoSQL data storing types and use cases, constructed using [13]

Type	**Examples**	**Use cases**
Key-value	Amazon Dynamo, Redis	Caching data from relational databases Storing configuration data
Document	JavaScript Object Notation (JSON)	Variety of data Large objects
Graph	Titan	Visualizations of relationship networks
Column	Cassandra	Large volumes of data High read and write performance High availability

MongoDB is a document-type database that stores data as binary encoded JSON like objects (BSON) that have no predefined schema [12]. According to a ranking of DB-Engines.com, MongoDB is the most popular NoSQL database [14]. MongoDB is aimed to take flexible data model, scalability and performance from the NoSQL databases without losing the query language, secondary indexes and consistency of RDBMS databases [15].



```
{
    "_id": 1,
    "name": "Adam",
    "age": 37,
    "occupation": "Lumberjack"
},
{
    "_id": 2,
    "name": "Bernie",
    "age": 42
},
{
    "_id": 3,
    "name": "Charlie",
    "age": 50,
    "occupation": "Taxi-driver"
}
```

Figure 1. MongoDB collection using the same data as Table 1, adapted from [16]

By comparing examples of MySQL and MongoDB using the same data, the most significant difference is in the second row of Table 1, and the second document of Figure 1, consisting of "_id", "name" and "age" fields. MySQL table has NULL value for the empty value of Occupation column of the Id 2 while MongoDB does not have Occupation key at all in the _id: 2 of Figure 1. The difference exists because MySQL table has a predefined structure of fields that need to be inserted to every row whether they are empty or not while MongoDB does not have a predefined structure for either the fields or their datatypes [15,17].

2.3 Frameworks

In addition to databases, some other tools are needed for a Big Data system [3]. Big Data frameworks are needed for example for collecting and categorizing the data, moving the data around the system and monitoring the data [3]. Hadoop and High-Performance Cluster Computing (HPCC) are used as examples in this section. Hadoop is chosen as an example because it is a popular open-source framework [18]. HPCC is chosen because it is an open-sourced challenger of Hadoop that is not under Apache Foundation [19].

2.3.1 Apache Hadoop

One of the best-known frameworks around Big Data is Hadoop. Hadoop consists of distributive file system Hadoop Distributed File System (HDFS) and a programming framework called Map Reduce [20]. Hadoop was released 2011 by Apache Foundation, which is known also from other open source projects such as web server software Apache and Java Servlet Container Tomcat [21]. Initial versions of Hadoop were created as early as 2006 by Doug Cutting [20]. Hadoop has been used by some of the world's largest companies, such as LinkedIn, Spotify and Twitter [18]. In the gaming industry, it has been praised by companies like Electronics Arts and Blizzard that have millions of players online playing their games around the clock [1,2].

The technologies used by Hadoop, HDFS and Map Reduce, are a lot different from relational databases and the structured queries used in them. The data inside HDFS is not structured, it is more like a huge cluster with all kinds of information spread widely around it. MapReduce is a tool that helps to find out the needed information inside of the chaotic environment. [20]

The principle of MapReduce is that it starts by mapping the messy information to help grouping it up [5]. After the grouping is done, the reduce phase starts leading to output information [5]. Word count example is a good method to show how MapReduce is practically working step-by-step.

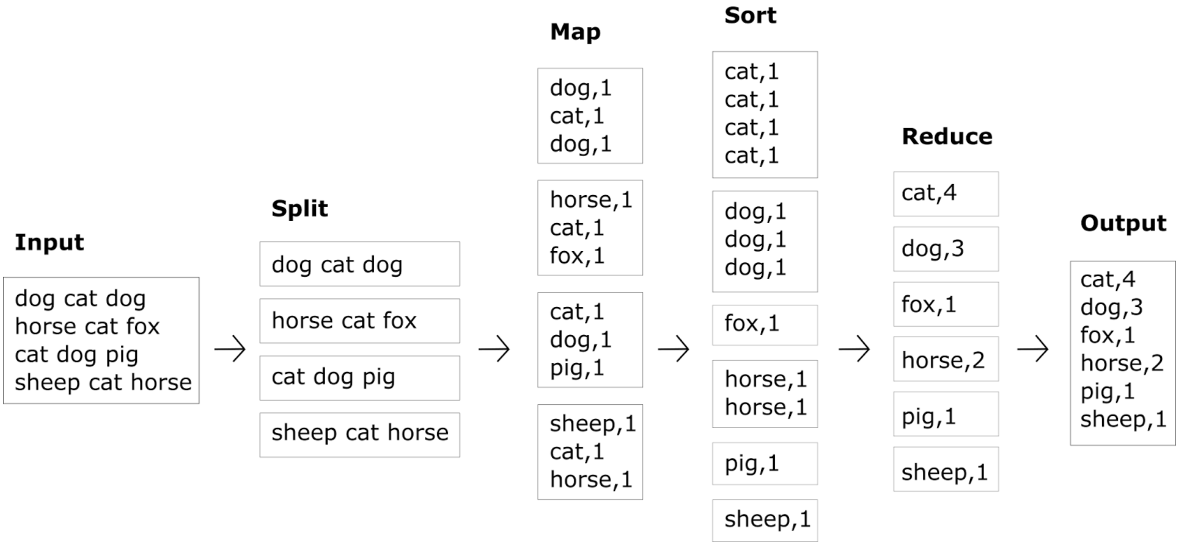


Figure 2. Wordcount example demonstrating MapReduce, adapted from [5]

2.3.2 High-Performance Cluster Computing

High-Performance Cluster Computing (HPCC) is an open-source computing platform that has been developed by LexisNexis [22]. The development of HPCC began 1999 and the first production stage applications appeared the following year. The project was not open-source until 2011 when LexisNexis decided to challenge Hadoop with their solution [19].

The main tool in HPCC is the Enterprise Control Language (ECL) programming language that is used similarly as Hadoop’s tools to refine unconstructed data to useful, constructed and sorted data [4]. LexisNexis claims that HPCC is faster than Hadoop, compiled C++ used by ECL is one of the key factors which uses mostly Java in its MapReduce operations [22]. ECL is used in The Data Refinery, known as Thor, and Data Delivery Engine, known as Roxie [4]. Thor is a similar tool to Hadoop’s MapReduce, it receives enormous amounts of data and sorts it for a more useful format [4]. Roxie on the other hand is a read-only processing cluster which is used for ultrafast database searches [4].

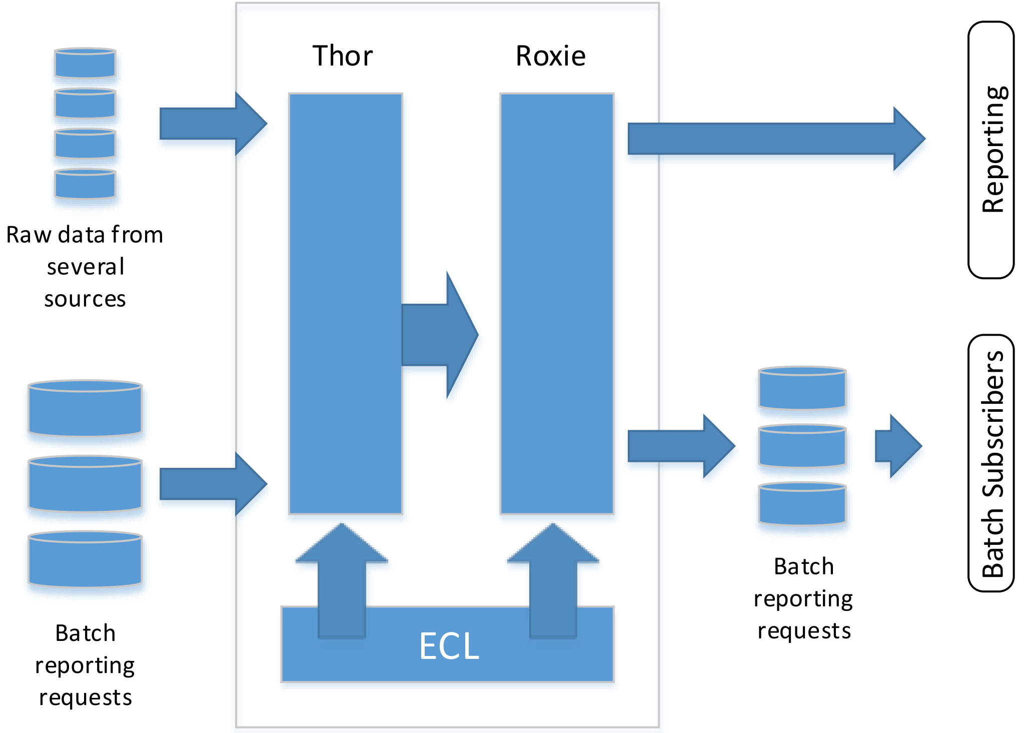


Figure 3. Thor, Roxie and ECL in the center of HPCC. Adapted from [4].

Older frameworks like Hadoop and HPCC have paved the way for newer alternatives like Apache's Spark and Flink, both of which have passed Hadoop in performance tests [23]. Different frameworks have different features and therefore comparing them is not easy. In the end of the day companies should choose the Big Data framework that is best suited to their needs and therefore Hadoop can still be a valid option even though it does not have the best performance.

3 EVOLUTION OF VIDEO GAMES

Video game industry has been affected a lot by Internet growing to current state of it during the nineties and the 21st century. Video games are no longer a privilege of gaming consoles, video games have become more social and nowadays some people playing video games are recognized as professional athletes [24].

3.1 Evolution of game types

Adding the Internet and the data aspect to games has changed the gaming sector completely. With the oldest gaming consoles, you could not imagine being in the same game session with someone who is not in the same room with you, let alone not in the same country with you. There were multiplayer games like Pong that included two players in the same game but you had to be there watching the same TV screen [25].

Eventually there emerged multiplayer games that could be played in the same network with different clients rather than just games with two controllers or “hot-seat” style of multiplayer mode. During the nineties people started to gather to have demo competitions and eventually to play multiplayer games at demo parties such as Assembly in Finland [26]. People carried their desktop computers to the venue and in addition to watching the demo competitions they played games like Doom (1993) and Quake (1996) against each other over the Local Area Network (LAN).

Proper Internet connections started to spread in the Western world during the beginning of 21st century [27]. Suddenly people could play even fast-paced first-person shooter games like Quake against each other without leaving their homes if they were connected to the Internet. Enthusiastic players no longer had to carry their whole computer setup to somewhere else to play with their friends.

Single-player games have also gotten more features with the addition of the Internet. For example, leaderboards can be held in real-time and you can even race against “ghost” players in some games to see how the record times and results have been achieved [28]. It is also possible to show statistics of other players’ in-game choices right after player has finished the game himself [29].

3.2 Evolution of developing video games

Both developing and playing the video games have changed a lot with the help of Internet and the possibility to move game data across the world quickly instead of just projecting the game from a machine to television. Before Internet grew big enough, games could not be updated easily after they were published. Therefore, developers had to make sure that everything worked as intended before publishing the game. If there were bugs in the game, they could be there forever. For some games, it was possible to do patching with floppy disks and later with Compact Disks but that meant sending the physical object through mail service [30].

Nowadays publishers usually have all their players connected to Internet so they can update the games whenever they want by publishing a so-called “patch” for players to download. They can even involve players in the testing process with open and closed beta tests. Even further, it is possible to collect money for developing the game by making a so-called Early Access release of the game [31]. Players are informed that the game is still in developing phase but they can already play some parts of it [31]. If any bugs occur during their playing experience, developers might get all the needed information about them automatically if they happen to collect the right data.

Testing takes a lot of resources to do properly and if the testing group is too small, skill levels between players might vary from beginner to professional player and that makes the balancing impossible. When there are thousands of players in a game that is not even released yet, it helps the developers a lot in the process of balancing. For example, a battle arena game Battlerite by Stunlock Studios was in Steam Early Access from September 2016 to the end of October 2017 and during that time it had more than 1000 active players every month [32].

3.3 Evolution of monetization models

In the gaming business, it was a standard for a long time that customer pays the full price of the game upfront and receives all the features of it right away [33]. Times have changed though and lots of payment models have entered the industry instead of the classic Premium model [33].

Table 3. Most common monetization models, constructed from [33]

Name	**Payment**	**Example**	**In-game purchases**
Premium	Upfront	Overwatch	Cosmetic
Free to play	Optional	Clash Royale	Helpful, boosting
Subscription	Once in a month	World of Warcraft	Cosmetic

Free to play (F2P) games such as Clash Royale and Clash of Clans have made huge profits for their developer Supercell during the last few years [34]. Players get in easily because the games are free and the gameplay is addictive [35]. In the beginning player feels like he is progressing well in the game but at some point, the progress gets slower and player starts to get offers for in-app purchases (IAP) [35]. Spending real money speeds things up in F2P games allowing players to make faster progress through ranks [33]. It is still good to keep in mind that unfair advantage should not be gained by spending money in the game. Player that spends no money in the game should have the same chances of winning in the game as a player who spends money or potential paying customers might quit playing the game [33]. For example, in Clash Royale players can either play a lot to gain in-game currency or they can spend real money for it and save time [35]. Players also gain loot boxes for winning in the game but they must wait for hours before they can open them unless they use the in-game currency for opening them [35].

For the games that are using premium and subscription models, developers need to build enough collective demand so that they can get enough people to buy their game and make the project profitable enough. The economic architecture for these Pay-to-play (P2P) games is Development-Monetization-Acquisition-Retention (D-M-A-R) while F2P economic architecture is Acquisition-Retention-Monetization-Development (A-R-M-D). This means that after the game is developed, it needs the monetization by getting the buying customers, acquisition by players playing it and retention meaning that players enjoy the game and keep playing it. If players feel like the game was not worth the money, it is highly unlikely that they would by a sequel for the game by the same gaming company. [33]

The differences between economic architectures of the F2P and P2P mean that while all the players of a P2P have invested money in the game, it is common that 90-99% of players in F2P game stay as free players while only 1-10% are paying players. Of course, some players might use money one month and then play several months for free so these percentages are not carved in stone. The goal for the developer is that at least some of the non-paying customers transform eventually to paying customers. Paying customer types from the least consuming to most consuming are called Minnows, Dolphins and Whales. [33]

3.4 Most common problems of gaming industry

Game development is like software development in terms of constantly finding and fixing bugs and defects in the product in addition to creating new features. A survey “What went wrong? A survey of problems in game development” (F. Petrillo, M. Pimenta, F. Trindade & C. Dietrich 2009) listed the occurrences of common problems in game development [36]. Table 4 introduces five most occurring problems in the study.

Table 4. Five most common problems in [36]

Problem	Occurrence	Description
Unrealistic scope	75%	Unrealistic and too ambitious expectations of how much developers can produce in the given time
Feature creep	75%	Adding features too late for the project to be developed and tested enough instead of using features that were originally planned
Cutting features	70%	The original plan has had too many features and functionalities and they need to be cut to maintain the schedule and the budget of the project
Design problems	65%	The original designs have not been good enough and they need to be modified during the development. There might be too specific design or not enough design.
Delays	65%	Release dates get pushed further because the original schedule has been planned overoptimistically

During these eight years after the survey, a lot has happened in the gaming industry and these answers could be very different today. However, looking at the game design of popular multiplayer games, it seems like problems still revolve a lot around the design and the number of features [37,38].

During the last decade, there have been published lots of Multiplayer Online Battle Arena (MOBA) games with more than ten or even more than hundred different playable characters that are playing against each other [39]. For the game developers, it becomes a continuous process of balancing; redefining the design of the heroes so that every character is good against some characters but at the same time it is possible to counter every character somehow. It is possible that all the characters have one primary skill and 1-4 different skills in addition to that [40]. If you have ten people playing against each other, in teams or individually, with everyone playing different characters, there might be more than 40 different skills that need to be useful but not overpowered. Failing the balancing process causes players to get frustrated and even to stop playing the game [41]. In the past the game development company might have already gotten their money at this point, but with the monetization models today it is important avoid losing active players [33].

While game developers try to find the bugs and the balance between characters, they need to find the balance in how they operate the monetization of their game. If the game is using the F2P monetization model, the game needs to be enjoyable for both paying and non-paying players. Similarly, to a wide pool of different characters suiting the needs of different people, there should be diverse pool of in-game purchases to choose from. When player has spent money in the game at least once it is more likely that he will stay playing the game in the future. Going too greedy with the monetization can cause a huge disaster in the public relations as Electronics Arts noticed with their Star Wars Battlefront 2 before the game was even launched [37].

When the games are getting more complex in terms of their features, testing and monetization, it is important to gather enough data and analyze everything properly before jumping to conclusions. There is only so much that a small group of game developers and testers can do if the game is designed to be played by millions of people with different skillsets and mindsets. As mentioned in the section 2, Big Data frameworks are useful for both collecting and analyzing data and therefore they can help with these problems.

4 USAGE OF BIG DATA IN VIDEO GAMES

There are a few examples of gaming companies that have implemented Big Data solutions successfully from the beginning of their lifespan and there are some companies that have adapted to it later. Gaming companies do not necessarily have to use Big Data to provide their services properly unlike giant social media services that handle ridiculous amounts of data every second of the day. Nevertheless, companies such as Supercell and Zynga, are fine examples of how much Big Data can help a gaming company to succeed. Petri Kärkäs from Supercell participated in Aalto Big Data Breakfast on December 2015 and in his presentation, he listed the following use cases for Big Data in his company [42].

- Daily metrics
- Game balance and player population metrics
- A/B testing
- Marketing optimization
- Bot detection

Supercell is not the only company using Big Data in its advantage with the use cases. This section introduces some of the use cases with examples. Bot detection is left out because not enough material was found linking Big Data for it.

4.1 Balancing

Balancing is an important part of developing both single-player and multiplayer games to keep the game fair and fun [43]. The difficulty of single-player games need to be increased gradually so that players will not quit because the game is too easy or too difficult [44]. Multiplayer games need balancing for asymmetrical features, such as player characters with different attributes and abilities [43]. Big Data can be used for balancing both single-player and multiplayer games but this section focuses on the multiplayer aspect.

4.1.1 Player character balancing

Blizzard Entertainment has gotten more than 35 million players to play their newest first-person shooter game Overwatch [45]. The game was launched on May 2016 with 22 different playable characters that are known as heroes, four different game modes and nine different maps. There are four different hero categories and all the heroes have different skills that Blizzard needs to have balanced [46]. Games with this amount of heroes and abilities cannot be completely balanced for all the game modes, maps and player skill levels from beginners to professional players. Therefore, some combinations of heroes become the so-called “meta”, meaning that these heroes and combinations are preferred instead of others.

During 2017 players were voicing concerns that certain, more mobile heroes have become too overpowered and that so-called “dive comp meta” does not allow playing other than these heroes and it makes the game boring [38]. Dive comp in short means using a composition of certain mobile heroes to make coordinated attacks against the enemy [38]. The Game Director Jeff Kaplan replied with a long post addressing many issues of the game at the same time with the following parts addressing the dive issue using statistics to show that everything is fine [47].

First, he addressed the six most played heroes from the more casual game mode, Quick Play, and only one of them, Genji, is one of the “dive comp meta” heroes [47]. Secondly, he listed the seven most played characters from the Competitive game mode and that list had both very mobile “dive comp” oriented heroes and slower heroes that are considered to get battered by “dive comp” heroes [47]. He ended his hero statistics by stating that the most played hero by the top third of the competitive players was Ana [47]. Ana is a support hero that was not considered to be the best option for dive comps rather than Lucio or Zenyatta [48]. Kaplan’s reply was an effective method to use statistics gained using Blizzard’s Big Data tools to prove a point to the player base and to avoid making hasty decisions only because the players felt that a change was needed.

4.1.2 Game environment balancing

In addition to balancing the heroes, some of the maps in the games need balancing too and Overwatch forums had a good example of this. In the game mode called “Assault” the attacking team consisting of six people has ten minutes to capture two capture points while defending team, also a roster of six people, does everything they can to prevent this [49].

Players started to voice concerns in the official game forum about the game mode favoring the defending side [50]. When a player posted a question about win rates in Assault maps in Quick Play game mode, Game Director Jeff Kaplan replied with the statistics from a time period of about two months [50].

Table 5. Win-percentages in Quick Play maps [50]

Map	Attacking side win-%	Defending side win-%
Hanamura	49.79%	50.21%
Horizon	54.50%	45.50%
Temple of Anubis	49.92%	50.08%
Volskaya	49.76%	50.24%
Eichenwalde	46.13%	53.87%
Hollywood	49.91%	50.09%
King's Row	50.14%	49.86%
Numbani	49.40%	50.60%
Dorado	49.66%	50.34%
Route 66	49.94%	50.06%
Watchpoint Gibraltar	50.13%	49.87%

In this list, there are all the Assault maps in the game: Hanamura, Horizon Lunar Colony (Horizon), Temple of Anubis and Volskaya Industries (Volskaya). As the data shows, the game mode seems balanced except for Horizon being attack-favored which is the opposite of what the player was thinking about Assault maps.

4.2 Monetization

Monetization is the most important factor for a game company because without a good monetization system it does not really matter whether the game is a hit or not. In the end of the day the main goal for a game company is to be a profitable company and without a well-designed monetization model that goal is difficult to achieve. Big Data can be very useful for the F2P monetization model because the economic architecture A-R-M-D allows to refine the game's monetization continuously using actual player data in the process [33]. Premium model does not necessarily have this advantage because most of the development is done before the monetization [33].

Zynga was one of the first companies to adopt the data-driven game developing in their F2P Facebook games such as FarmVille and Mafia Wars. Founding member Andrew Trader said in an interview with The Wharton School: "*The dirty little secret of Zynga is, of the five corporate values, none more important than being metrics driven. For Zynga, that meant if you can't measure something, don't build it.*". Zynga followed closely their reach, retention and revenue with the retention being the most important factor. Their revenue was gained by selling the in-game items their customers wanted. As a rough example, it was mentioned that women bought things mostly for the looks while men bought things to beat their friends in the game. They used their metrics to find out what their players wanted and delivered it. [51]

By looking at the player numbers and sales figures, Supercell is one of the best examples using F2P model and Big Data to get the maximum profit without scaring the non-paying players away. Supercell uses Amazon's cloud services, particularly Amazon MapReduce along with a Hadoop framework to gather and analyze more than 45 billion in-game events per day [52]. Basically, every action happening in the games of Supercell are sent to their cloud and analyzed using a large dataset to make sure that everything is under control [52].

4.3 Testing

There are various methods and tools for testing video games. Typically, game testers and Quality Assurance (QA) try to make sure that [53]:

- The game is balanced
- The game does not crash in any situations
- Both old and new features are working as intended

Big Data offers help for game testing by giving option to gather large samples of test data and to analyze the data live.

"*Our main business challenge is to figure out what makes people play our games, what makes them fun. We are creative first; then, we use the data to validate those decisions. We have a hypothesis, the game goes live, and we test our hypothesis.*" -Janne Peltola, data scientist of Supercell in a case study of Hewlett Packard Enterprise about the big data solutions of Supercell. [54]

A/B testing means that in addition to the live version of the game there is also a different variation of the game. These variations can include for example modifications in UI (User Interface) or abilities in player characters. The goal is to gather enough data with both the live version, also known as "control" version, and the variant version to see if the modification leads to desired results in the variant version. [55]

"*One interesting A/B test Supercell conducted was about Facebook connectivity. The test was designed to look at whether people who liked the games also wanted to encourage their friends to play, and to see how those variables affected player retention. In terms of A/B testing, we now pull the data into HPE Vertica. Something that took two to three hours in the past now takes four minutes*." [54]

Blizzard has a separate PTR (Public Test Realm) environment where all the players can test the next patch prior to actual release to live servers [56]. PTR is useful for testing new heroes and maps to make sure there are nothing game-breaking in them before they are released to millions of players [56]. Game-breaking things include for example overpowered abilities, overpowered player characters and bugs in maps or heroes that crash the whole game server. The problem with the testing in PTR is that it is not very popular compared to the live servers and therefore the amount of gathered data is really small compared to their live servers [57]. According to a forum post of the Game Director Jeff Kaplan in January 2017, only 0.26% of the Overwatch players have entered PTR so far [57]. At the time, the overall player count for Overwatch was somewhere between 20 and 25 million so player count in the PTR realm was somewhere between 52000 and 65000 [58].

Even though the PTR is not as popular as Blizzard would like it to be, more than 50000 players in a test environment providing data for them is still a lot more than the employee count, 9600, of the entire organization of Activision Blizzard [59]. It is also worth to mention that some players do not just play on the PTR but also report their findings to the official game forum or other Overwatch forums such as different Overwatch subreddits in the Reddit community [60].

5 DISCUSSION

It is impossible to know how big the video gaming industry would have gotten without the Internet and social aspects it has brought to video games. Without the help of Internet, gaming companies surely would not invest as much to multiplayer games. There probably would still be multiplayer games available but multiplayer would be only as addition to a single-player campaign. Free-to-play and Big Data would be bizarre concepts since neither would make any sense without the connections of today.

Both the gaming industry and Big Data have been studied a lot probably because both have become so big commercially that they simply cannot be ignored. Both can be easily studied separately but the actual topic of Big Data in Video Games is more difficult to analyze. Most of the sources available are case studies and whitepapers made by companies providing Big Data services for the gaming companies. It does not mean that these sources cannot be trusted but it is good to remember that while these studies state a lot of facts about the benefits of Big Data for gaming companies, they are also good marketing material for the Big Data service providers.

Big Data is still a relatively new thing and it surely has not reached its full potential in the video game industry. Even though there are successful early adopters such as Supercell and Zynga, it is possible that Big Data and data-driven development are not seen as needed solutions for game developing companies. Of course, it is possible that there are many who have tried and failed without anyone noticing. Whatever is the case, games will keep evolving with the possibilities given by the technology like they have done to this day.

6 CONCLUSIONS

Video game industry has grown bigger than anyone could possibly imagine when the first commercial video games were published. This would not happen without the help of Internet even though games had their fair share of players even before proper Internet connections spread all around the world. Video games are played everywhere with all kinds of devices and internet connection is taken for granted for many of these new game types that have emerged in the 21st century.

The amount of data in the Internet has grown unexpectedly fast and eventually the traditional relational databases could not keep up anymore with the volume, velocity and variety of data. While relational databases still serve their purpose for smaller volumes, velocities and varieties of data, the concept of Big Data was needed to handle the challenge with NoSQL databases and frameworks like Hadoop and HPCC.

Big Data has brought more to the table of game developers in terms of balancing, monetization and testing. Decisions for balancing and monetization can still be made with hunch but there is always data available to study the effects properly. Developers can also reason their actions or lack of action by showing raw statistics to their players. Hopefully players will still be heard even though developers can just look at the data and see if something seems to be wrong instead of reading the players' opinions in the game forums. After all, it does not matter whether the game is good, which monetization model is used or how accurately the statistics can be pulled if there are not enough paying customers for the game.

REFERENCE

1. R. Taneja, ”Video Games: The Biggest Big Data Challenge”, presented at the O’Reilly Strata conference, Santa Clara, CA, February 27, 2013.

2. C. Burkhart and J. Irwin, Building a Near Real-Time Pipeline for All Things Blizzard, presented at Elastic{on} 2017, San Francisco, CA, March 8, 2017.

3. Michael Frampton, ”Big Data Made Easy”, SpringerLink, 2015, pp. 1-3.

4. B. Furht and F. Villanustre, Big Data Technologies and Applications, SpringerLink, 2016, pp. 3, 164-171.

5. G. Harrison, ”Next Generation Databases - NoSQL, NewSQL, and Big Data”, SpringerLink, 2015, pp. 4-19.

6. “Big Data and RDBMS: Can they coexist?”. [Online]. Available: https://www.informationweek.com/big-data/big-data-and-rdbms-can-they-coexist/d/d-id/1324939? [Accessed: 28-Nov-2017].

7. “CMS Usage Statistics”. [Online]. Available: https://trends.builtwith.com/cms [Accessed: 28-Nov-2017].

8. “Database Description”. [Online]. Available: https://codex.wordpress.org/Database_Description [Accessed: 28-Nov-2017].

9. ”General Information”. [Online]. Available: https://dev.mysql.com/doc/refman/5.7/en/introduction.html [Accessed: 05-Dec-2017].

10. “Examples of Common Queries”. [Online]. Available: https://dev.mysql.com/doc/refman/5.7/en/examples.html [Accessed: 29-Nov-2017].

11. “Big Data Databases Explained”. [Online]. Available: http://basho.com/resources/big-data-databases/ [Accessed 29-Nov-2017].

12. C. Gyrödi, R. Gyrödi, G. Perleche and A. Olah, ”A Comparative Study: MongoDB vs. MySQL”, 13th International Conference on Engineering of Modern Electric Systems (EMES), 2015.

13. “Types of NoSQL databases and key criteria for choosing them”. [Online]. Available: http://searchdatamanagement.techtarget.com/feature/Key-criteria-for-choosing-different-types-of-NoSQL-databases [Accessed: 29-Nov-2017].

14. ”DB-Engines ranking”. [Online] Available: https://db-engines.com/en/ranking [Accessed: 02-Dec-2017].

15. ”MongoDB Architecture”. [Online]. Available: https://www.mongodb.com/mongodb-architecture [Accessed: 02-December-2017].

16. ”The bios Example Collection”. [Online]. Available: https://docs.mongodb.com/manual/reference/bios-example-collection/ [Accessed: 02-December-2017].

17. ”INSERT Syntax”. [Online]. Available: https://dev.mysql.com/doc/refman/5.7/en/insert.html [Accessed: 02-Dec-2017].

18. “Powered by Apache Hadoop”. [Online]. Available: https://wiki.apache.org/hadoop/PoweredBy [Accessed 29-Nov-2017].

19. “LexisNexis Will Open-Source Its Hadoop Alternative for Handling Big Data”. [Online]. Available: http://readwrite.com/2011/06/15/lexisnexis-open-sources-its-hadoop-alternative/ [Accessed: 29-Nov-2017].

20. M. Chen, S. Mao, Y. Zhang and V.C.M. Leung, ”Big Data - Related Technologies, Challenges and Future Prospects”, Springer Briefs in Computer Science, 2014, pp. 16.

21. ”Apache Project List”. [Online]. Available: https://www.apache.org/index.html#projects-list [Accessed: 30-Nov-2017].

22. “HPCC Systems: Data Intensive Supercomputing Solutions”. [Online]. Available: http://cdn.hpccsystems.com/whitepapers/wp_data_intensive_computing_solutions.pdf [Accessed: 29-Nov-2017].

23. J.Veiga, R.R.Expósito, X.C.Pardo, G.L.Taboada, J.Tourifio: ”Performance evaluation of big data frameworks for large-scale data analytics”, 2016 IEEE International Conference on Big Data (Big Data), IEEE 2016.

24. ”The U.S. Now Recognizes eSports Players As Professional Athletes”. [Online]. Available: https://www.forbes.com/sites/insertcoin/2013/07/14/the-u-s-now-recognizes-esports-players-as-professional-athletes/ [Accessed: 06-Dec-2017].

25. ”Pong Game”. [Online]. Available: http://www.ponggame.org/ [Accessed: 29-Nov-2017].

26. ”Historia – Assembly”. [Online]. Available: http://www.assembly.org/summer16/about-us/history/ [Accessed: 29-Nov-2017].

27. ”Our World in Data”. [Online]. Available: https://ourworldindata.org/internet/ [Accessed: 29-Nov-2017].

28. ”The usage of Ghosts in Forever (Tutorial – part 1)”. [Online]. Available: https://tmunited.wordpress.com/2008/05/05/the-usage-of-ghosts-in-forever-tutorial-part-1/ [Accessed: 29-Nov-2017].

29. ”Telltale Games: The majority of The Walking Dead players try to do the right thing”. [Online]. Available: https://venturebeat.com/2012/08/15/telltale-games-the-walking-dead-statistics-trailer/ [Accessed: 29-Nov-2017].

30. ”Sierra Patch Disk”. [Online]. Available: http://i.imgur.com/GO26lHl.jpg [Accessed: 29-Nov-2017].

31. ”Early Access Games”. [Online]. Available: http://store.steampowered.com/earlyaccessfaq/ [Accessed: 29-Nov-2017].

32. ”Steamcharts – An ongoing analysis of Steam’s concurrent players”. [Online]. Available: http://steamcharts.com/app/504370#All [Accessed: 29-Nov-2017].

33. M. Dadidovici-Nora, Paid and Free Digital Business Models - Innovations in the Video Game Industry, Digital Economic Journal, No. 94, 2014.

34. “Clash of Clans” Maker Supercell Posts $2.3B In Revenue, $930M In Profit For 2015 As Growth Slows. [Online]. Available: http://www.ibtimes.com/clash-clans-maker-supercell-posts-23b-revenue-930m-profit-2015-growth-slows-2333237 [Accessed: 29-Nov-2017].

35. ”Clash Royale – Deconstructing Supercell’s Next Billion Dollar Game”. [Online]. Available: https://www.deconstructoroffun.com/blog//2016/02/clash-royale-next-billion-dollar-game.html [Accessed: 29-Nov-2017].

36. F.Petrillo, M.Pimenta,, F.Trindade and C.Dietrich, ”What went wrong? A survey of problems in game development”, Computers in Entertainment (CIE) - SPECIAL ISSUE: Media Arts and Games, Vol.7 Issue 1, 2009.

37. ”EA Is Now Ironically Stuck With $60 ’Battlefront 2’ And No Good Way To Re-Monetize It”. [Online]. Available: https://www.forbes.com/sites/insertcoin/2017/11/24/ea-is-now-ironically-stuck-with-60-battlefront-2-and-no-good-way-to-re-monetize-it/ [Accessed: 29-Nov-2017].

38. ”The Problem With Overwatch’s ‘Dive’ Meta”. [Online]. Available: https://compete.kotaku.com/the-problem-with-overwatchs-dive-meta-1796828151 [Accessed: 29-Nov-2017].

39. ”MOBA Reviews”. [Online]. Available: https://mmos.com/review/moba [Accessed: 29-Nov-2017].

40. ”DOTA 2 Heroes”. [Online]. Available: http://www.dota2.com/heroes/ [Accessed: 29-Nov-2017].

41. ”Applying Risk Analysis To Play-Balance RPGs”. [Online]. Available: https://www.gamasutra.com/view/feature/131252/applying_risk_analysis_to_.php[Accessed: 30-Nov-2017].

42. P. Kärkäs, ”Data Science and Big Data at Supercell”, presented at Aalto Big Data Breakfast, Espoo, Finland, December 12, 2015.

43. ”Understanding Balance in Video Games” [Online]. Available: https://www.gamasutra.com/view/feature/134768/understanding_balance_in_video_.php [Accessed 30-Nov-2017].

44. ”16 reasons why Players Are Leaving Your Game”. [Online] Available: https://gameanalytics.com/blog/16-reasons-players-leaving-game.html [Accessed: 30-Nov-2017].

45. Tweet by PlayOverwatch. [Online]. Available: https://twitter.com/PlayOverwatch/status/919925924769906688 [Accessed: 3-December-2017].

46. ”Heroes – Overwatch”. [Online]. Available: https://playoverwatch.com/en-us/heroes/ [Accessed: 29-Nov-2017].

47. “OW Update/Balancing cycle is excruciatingly slow: Part 2”. [Online]. Available: https://us.battle.net/forums/en/overwatch/topic/20757706588?page=2#post-27 [Accessed: 29-Nov-2017].

48. ”Whose dive is it anyway?”. [Online]. Available: https://www.overbuff.com/blog/2017-07-02-whose-dive-is-it-anyway [Accessed: 29-Nov-2017].

49. ”Game Overview”. [Online]. Available: https://playoverwatch.com/en-us/game/overview/ [Accessed: 29-Nov-2017].

50. “2 CP win rates in Quickplay, Blizzard?”. [Online]. Available: https://us.battle.net/forums/en/overwatch/topic/20757527371 [Accessed: 29-Nov-2017].

51. “From Virtual Barnyards to Real Dollars: Andrew Trader on Zynga, ‘Gamification’ and the Power of Analytics”. [Online]. Available: http://knowledge.wharton.upenn.edu/article/from-virtual-barnyards-to-real-dollars-andrew-trader-on-zynga-gamification-and-the-power-of-analytics/ [Accessed: 29-Nov-2017].

52. ”Supercell Case Study”. [Online]. Available: https://aws.amazon.com/solutions/case-studies/supercell/ [Accessed: 29-Nov-2017].

53. L. Levy and J.Novak, ”Game Development Essentials – Game QA & Testing”, 2009, pp. 58-69.

54. ”Supercell adopts HPE Vertica Analytics Platform”. [Online]. Available: https://www.vertica.com/wp-content/uploads/2017/06/Supercell-Success-Story.pdf [Accessed: 29-Nov-2017].

55. K. Watanabe, T.Fukamachi, N.Ubayashi and Y.Kamei, Poster: Automated A/B Testing with Declarative Variability Expressions, IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE 2017.

56. ”Public Test Region”. Online. Available: http://overwatch.wikia.com/wiki/Public_Test_Region [Accessed: 29-Nov-2017].

57. “Is the Sky Really Falling?”. [Online]. Available: https://us.battle.net/forums/en/overwatch/topic/20752500232?page=5#post-99 [Accessed: 29-Nov-2017].

58. Tweet by PlayOverwatch. [Online]. https://twitter.com/PlayOverwatch/status/824757676693270529 [Accessed: 29-Nov-2017].

59. ”Activision Blizzard on Forbes Global 2000”. [Online]. Available: https://www.forbes.com/companies/activision-blizzard/ [Accessed: 29-Nov-2017].

60. ”Undocumented PTR nerfs to Winston and Reinhardt”. [Online]. Available: https://www.reddit.com/r/Competitiveoverwatch/comments/7ei1c4/undocumented_ptr_nerfs_to_winston_and_reinhardt/ [Accessed: 29-Nov-2017].