POSSIBILITIES AND CHALLENGES OF ADOPTING BIG DATA ANALYTICS AT CCP

Research report

Spring 2012

Kristófer Hannesson
B.Sc. in Computer Science

Confidential until June 1, 2015

Author: Kristofer Hannesson
Kennitala: 200882-4589

Supervisor: Björn Þór Jónsson
Examiner: Yngvi Björnsson

T-619-LOKA
School of Computer Science

Abstract

Businesses today are both generating ever more granular data than before and have access to abundant new data sources. This has resulted in them working with orders of magnitude larger and more complicated data sets than before. These large and often unstructured data sets are referred to as "big data" and powerful technologies have been created in recent years to help businesses store and analyse them. In many cases big data analysis, led by Data Scientists, has become central to businesses' innovation and decision making.

The game developer and publisher CCP hf is taking its first steps towards adopting big data analytical capabilities. CCP operates in the highly competitive market of massively multiplayer online games. Their main game is the subscription based EVE Online with around 400 thousand active players. CCP will release Dust 514 later this year, a game with a different business plan and a potential player base is in the millions. CCP has decided to modify its strategy to be more data-driven and intends to adopt big data technology and analysis to help manage further development of their games and business. This will require CCP to mature and change a lot of their analytical pipelines and processes as well as their general attitude towards data use.

Útdráttur

Fyrirtækjarekstur í dag bæði skapar nákvæmari gögn en áður og hefur aðgang að mörgum nýjum gagnauppsprettum. Þetta hefur leitt til þess að þau vinna með margfalt stærri og flóknari gagnasöfn en áður. Þessi stóru og oft óskipulegu gagnasöfn eru kölluð „big data“ og öflug tækni hefur verið þróuð undanfarin ár til að hjálpa fyrirtækjum að geyma og greina þau. Í mörgum tilfellum hefur big data gagnagreining, leidd gagnavísindamönnum, orðið þungamiðja nýsköpunar og ákvarðanatöku fyrirtækja.

Leikjaframleiðandinn CCP hf er um þessar mundir að taka fyrstu skref í átt þess að taka upp big data gagnagreiningar. CCP starfar í hinum mjög samkeppnisharða markaði fjölspilunarleikja á netinu. Aðal leikur þeirra er áskriftarleikurinn EVE Online sem hefur um 400 þúsund virka spilendur. CCP mun gefa út leikinn Dust 514 seinna á þessu ári. Um er að ræða leik sem hefur nýtt viðskiptamódel og líkur eru á að fjöldi spilara hans hlaupi á milljónum. CCP hefur ákveðið að breyta rekstri sínum á þann veg að stjórnast í auknum mæli af göngum og ætlar að tileinka sér big data tækni og greiningar til að aðstoða við leikjaþróun þeirra og viðskiptamódel. Þessi ákvörðun krefst þess að CCP þroski og breyti mörgum sinna greiningaraðferða sem og viðhorfi til notkunar gagna.

Acknowledgements

This research project would not have been possible without the support of many people. I wish to express my sincere gratitude to my supervisor, Professor Björn Þór Jónsson, for his excellent advice, patience, and support. Deepest gratitude are also due to the members of the supervisory committee, Associate Professor Hrafn Loftsson and Associate Professor Yngvi Björnsson, for their generous advice.

My employer, CCP hf, and my boss, Dr. Eyjólfur Guðmundsson, for making this project possible and for his valuable input.

All the interviewees at CCP for taking time out of their busy schedule to meet with me.

Finally, I am forever indebted to my wife, Akiko, my parents, and my friends for their understanding, endless patience and encouragement when it was most required.

Contents

List of Figures

List of Tables

1 Introduction

Many businesses today are faced with the challenge of handling much greater amounts of data than in the past. They need to combine all of their data sources in one data storage, extract valuable information from the data, and execute on the results. The sheer amount of data and its often unstructured nature makes it infeasible or even impossible to use traditional database management systems to store and manipulate it. This amount of data has given rise to the term "big data" and the associated challenges have resulted in new technologies created specifically to enable storage and manipulation of huge and unstructured data sets. In today's aggressive markets, the way companies make use of big data technology will have a deciding influence on whether they can stay competitive and gain an advantage. Technological solutions such as Hadoop solve the storage and manipulation problems and with that the challenge shifts to one of analysis and execution.

Developers and publishers in the gaming industry, in particular those who develop massively multiplayer online games (MMOs), depend on the ability of their games to attract and maintain a strong user base of players willing to exchange money for meaningful gaming experiences. These companies collect significant amounts data on their players and their conduct. They need to extract information from this data to help them understand player behaviour, improve the player experience, attract new and old players, and retain those already playing. A new highly competitive service model called free-to-play (F2P) has emerged and taken over much of the MMO market in recent years. In the P4F model players can play the MMO for free, compared to the traditional model of a monthly subscription fee, but are given the option to purchase items or services that improve or expand on the gaming experience. With this business model the analytical requirements have increased considerably as a smooth gaming experience leading to item purchases is critical.

This paper focuses on what big data is and what it means for businesses, and the challenges and opportunities big data adoption poses for Icelandic MMO developer and publisher CCP hf which currently runs a subscription based MMO and will later this year release a F2P game. The paper begins in section two with a discussion about big data, its origins, usage, and important considerations for businesses adopting it. In section three the big data platform Hadoop is introduced and its various aspects and capabilities explained. Section four covers CCP and its games EVE Online and Dust 514, along with CCPs current data analysis efforts. Section five introduces the results of detailed interviews conducted within CCP on why the company is adopting big data technology and analysis, the perceived challenges, and what they aim to achieve. That chapter ends with a discussion on the interview results and how CCP should react. The paper concludes in section six.

2 Big Data

The term big data refers to data sets which are usually so large that using traditional database management systems to store and extract information from them is infeasible or even impossible. These data sets can also be somewhat unstructured which, coupled with their size, presents unique storage, extraction and analysis challenges. These problems have become more pronounced over the last few years as the cost of data storage has decreased fast, but data access speed has not kept up. Entities in both the public and private sector are now storing more data about their operations and customer behaviour than ever before. This data contains value that is crucial for them to extract in order to improve or simply maintain competitiveness in the market where it can help them remain a step ahead of the competition. But it is not enough to just look at the local data, there is also a wealth of data available from outside the boundaries of the company that can have a significant impact on its business and therefore should also be incorporated into the analysis. The challenge today is not where to find the data, it is to figure out what to do with it—how to combine all the data sources, extract the value contained within, and execute on the results.

2.1 The Roots of Big Data

The amount of data created in the world has been increasing by an estimated 50% every year and while the available storage has been getting cheaper it has not managed to keep up.[8][5] This increase can in large part be attributed to more of the same data being available in greater detail, but increasingly it is due to the availability of entirely new sources of data such as sensors, videos, websites, social networks, public records, and so forth.[7] With all this new data the struggle became how to store it all, access it, analyse it and execute on the results. In past years the data may have been kept in separate and entirely different systems where communication between them was hard or even impossible. This was deemed no longer acceptable so the challenge became how to link all these sources of data into a "single version of the truth" which could then be mined, for example to identify trends, improve efficiency, and create more reliable forecasts.[1]

For Chief Intelligence Officers (CIO) it became clear that they needed a new platform to achieve a single storage and analysis solution that could accommodate both the sheer amount of data and its often unstructured nature. Existing databases and business intelligence solutions were not designed to cope with these new challenges of scale and structure.[12] Traditional Relational Database Management Systems (RDBMS) do not effectively scale to this size and attempting so increases management costs considerably, and strict schema definitions hinder work with unstructured sources of data where you may not know what is important until after the logging event when careful analysis of the data has been performed.[12] CIOs today need a system where the storage problem is solved, accuracy ensured (error free data), data sources linked together, and the whole data set can be crunched for information in a respectable amount of time. A comparison of aspects of traditional RDBMSs and these new business requirements is presented in Table 1.

	Traditional RDBMS	Big Data Needs
Data size	Terabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

Table 1: Comparion of RDBMSs and big data needs. Table taken from Hadoop: The Definitive Guide [19, p. 5]

These needs and challenges have brought about both the name "big data" and the technical advances that have produced systems capable of solving the above mentioned problems. The two most important technical aspects are MapReduce, a method that allows for the distribution of work and combination of results, and distributed file systems. These are discussed in detail section 3.

2.2 How Big Data Is Used

In May 2011 the McKinsey Global Institute released a report where they outline five ways in which using big data can create value [10]:

1. It can unlock significant value by making information transparent and usable at much higher frequency.
2. Organizations can collect more accurate and detailed performance information on all aspects of their business to make better management decisions.
3. It allows ever narrower segmentation of customers and therefore much more precisely tailored products or services.
4. Sophisticated analytics can substantially improve decision-making.
5. It can be used to improve the development of the next generation of products and services.

The McKinsey report goes on to stress that the way companies make use of big data technology will be a key factor in their competitive advantage and therefore instrumental to their growth. The ability to extract value from ever growing and more granular data sets is of high importance for continued operation in the market space, central for innovation, and absolutely crucial for new entrants. The company that is the first to spot a trend or discover a need can be the first to market and capture the majority of the audience. The company that understands big data and how to leverage it can succeed where others could not.

It perhaps comes as no surprise that the companies most heavily involved with the internet and digital services seem to be the most prominent users of big data technology. This technology has become absolutely central to the operations of entities such as Facebook, Yahoo, Google, eBay, Twitter and many other web-service

based companies. These companies gather massive amounts of data on their users and their behaviour such as "their likes and dislikes, their relationships with others and even where they are at any particular moment", but they do not share the details of their methods or their findings because they lie at the heart of their competitive advantage.[4] Large teams of people in these companies have spent significant resources on building technology and optimizing the behavioural analysis that drives their monetization strategy.

The following are examples of how some companies have utilized big data technology in their business. Many of them mention the Hadoop platform which is discussed in section 3.

2.2.1 Facebook

In early 2010 Ashish Thusoo, then Facebook's Engineering Manager of Data Infrastructure, said that the average Hadoop system usage per day amounted to scanning 135 TB of compressed data, running over 7500 jobs, and 80 thousand computing hours. Thusoo added that 200 people per month are behind these jobs and that through simplified systems such as Hive (discussed later) analysts and other non-programmers could easily create and run jobs in Hadoop. He described the types of applications as reporting, ad hoc analysis, machine learning, and many others.[16] By the middle of 2010 the Hadoop warehousing solution at Facebook was storing "15PB of data (2.5PB after compression) and [loading] more than 60TB of new data (10TB after compression) every day."[15] As an example of useful analysis, Facebook found that "the best single predictor of whether members would contribute to the site was seeing that their friends had been active on it", so Facebook made changes to provide members with information on what their friends had been doing and thus increased participation.[4]

But Hadoop is not limited to only batch-processing analysis because it also handles some real-time usage at Facebook such as Facebook Messaging, which combines e-mail, chat, and SMS; Facebook Insights, which provides real-time analytics across websites using Facebook plugins, Facebook Pages and Facebook Ads; and the Facebook Metric System, which is continually fed statistics from all hardware and software used at Facebook.[3]

2.2.2 eBay

In 2010 eBay built a 530 server Hadoop cluster which had grown five-fold by the end of 2011, at which point it was handling a variety of tasks such as inventory data analysis and using online behaviour to construct customer profiles.[12] Bob Page, VP of the Analytics Platform, stated that eBay got such "tremendous value" out of the cluster that they now "depend on it to run eBay."[12] eBay has made many adjustments to its website and services based on analysis on "bidding behaviour, pricing trends, search terms and the length of time users look at a page."[4] Each product category at eBay is an actively managed "micro-economy", an example of which is that when an expensive item gets many searches but few purchases, eBay

will find an insurance partner for the seller to increase sales.[4]

2.2.3 Cablecom (Swiss telecoms operator)

Cablecom crunched information on their customer behaviour and managed to reduce customer defection from 20% a year to less than 5%. They found that while most customers left in the 13th month they actually made the decision to leave as early as the 9th month, so Cablecom offered special deals to customers seven months into their subscription.[1]

2.3 Using Big Data Effectively

Big data technology can open up easy access for almost anyone to mountains of data. This accessibility in itself can be very beneficial because it can get more people interested in the data and help them answer some of the specific questions they may have about the business. But simple interest will soon get overwhelmed when the data analysis reaches certain levels of complication. Erik Brynjolfsson, director of the Massachusetts Institute of Technology's Center for Digital Business, stated that the big problem facing businesses today is "the ability of humans to use, analyze and make sense of the data."[7] This is where the key challenges and key opportunities with big data lie. Storing petabytes of data is of no use unless it is easily accessible and can provide managers and executives with clearly understandable results that guide them towards better business decisions. This is what big data needs to be, a tool that provides actionable metrics—not only information about how we were doing in the past or how we are doing right now, but information that helps answer the most difficult question facing a business: "What do we do next?" The technology is already in place, the challenge is putting it into the hands of the people who can derive the right information from it.

2.3.1 Data Science

The people who are responsible for extracting from data the information leading to important decisions must have solid mathematical and statistical abilities. IBM researcher Daniel Gruhl, who mined medical data and managed to find ways to improve treatment, stated that "the key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd ... [a]nd that makes it easier for humans to do what they are good at—explain those anomalies."[7] The combination of massive amounts of data and applied mathematics has lessened the need to construct behavioural models because the available data allows researchers to simply measure, track and present the behaviour.[2]

In early 2010 the industry of big data in the US was valued at over $100 billion and estimated to be growing by 10% each year, double the growth rate of the entire software business.[5] At the heart of big data usage are the above mentioned professionals who have the mathematical and computing skills to query massive amounts of data, analyse the results, and extract actionable information. The growth of the

industry has increased the need for these kinds of professionals and it is predicted that by 2018 the US alone will need additional 140-190 thousand people with "deep analytical skills" and "1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."[10]

The growth of the big data industry and the need for people with a combination of analytical, programming, and artistic capabilities (for result presentation) has heralded the rise of a new kind of professional, the "data scientist". In an article in The Economist on the need for such professionals, the data scientist was described as an individual who combines these skills "to extract the nuggets of gold hidden under mountains of data."[5] Facebook is among the companies that have created a data science group, and Jeff Hammerbacher, then Data Manager at Facebook, described it as follows:

> "[...] on any given day, a team member could author a multi-stage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization".[9]

Data science is not about simply warehousing data, performing random searches, and going with gut feelings. Data Science requires the creation and testing of hypotheses and verification that the conclusions drawn from the data are valid. According to DJ Patil, then Head of Data Products and Chief Scientist at LinkedIn, "the best data scientists tend to be 'hard scientists', particularly physicists, rather than computer science majors."[9] The reason given is that physicists have a combination of strong mathematical and programming abilities, and come from a division of science where their work depends on extracting as much as possible from the data they have. But perhaps the most valuable skill, the one that the underlying data crunching and statistical analysis work relies on, is the ability to communicate the results in an easily understandable and informative way. The managers and executives may not have the time or mathematical understanding to read long statistical analyses and may therefore miss out on the opportunities the data presents. Results of careful analysis are nothing if they cannot be communicated effectively to the managers and executives.

2.3.2 Actionable Information

Big data capabilities coupled with data science should lead to data-driven discovery and decision making on all levels, but most importantly on the managerial level. The data is full of noise and it is the job of the data scientists to find "useful, relevant, timely information and wisdom that we can act upon to help us reach our goals."[13] Data science needs to provide us with early-warning signals or emerging opportunities, to move us from "information to insight to action at the time it matters."[13] Older warehoused data is good for historical analysis, but the big players are moving more aggressively towards converting real-time data streams to real-time information streams, and that should be the eventual goal of every business that takes the step to adopt big data.

Big data analysis can provide actionable information that for example:

1. Helps you win customers from the competition
2. Helps you prevent customers from leaving
3. Helps you ensure customers come back
4. Improves customer loyalty
5. Improves your products
6. Increases customer use of your product
7. Increases productivity
8. Allows for customized marketing
9. Minimizes risk
10. Identifies opportunities
11. Provides predictions

Managers in business and government need to be open and accepting of this kind of governance which moves away from the managers making gut-decisions based on limited data, to making decisions supported by the actionable information the algorithms and data scientists provide. When the decision has been made, no matter the lead up, the follow-up should be careful analysis of its effects. This creates the previously mentioned need for managers capable of using the statistical analysis of big data to help with decision making.

2.4 Privacy and False Interpretation

With individuals leaving an ever larger digital-footprint there is cause for alarm over how this sometimes deeply personal information is being used. As previously mentioned, entities such as Facebook, Yahoo, Google, Facebook, eBay and Twitter gather massive amounts of data on their users behaviour, likes and dislikes, relationships, and where they like to go, but they do not talk much about exactly what they store or how they use it. As a response to privacy considerations Google, for example, provides users with the option to completely erase their search history and thus a significant part of their digital-footprint. Digital service providers are likely to come under ever more scrutiny as the data they collect on users and their behaviour becomes larger and more detailed. But putting it in another light, users are the key to the feedback loop that on one hand is intended to provide the service providers with more revenue, be it from targeted advertisements or increased customer spending, but on the other hand intended to provide better and even tailor-made service to the users. The bottom line, however, is that individuals could be identified from their digital-footprint and with it reveal more than they would like, and the question thus becomes whether or not they consent to the storage and use of this information. The ethical considerations of running this kind of business are beyond the scope of this paper, but they are certainly something that governments, businesses and customers need to carefully navigate to arrive at a satisfactory solution.

There is also another perhaps less emphasized risk factor when it comes to big data analytics: false discoveries. Data sets can be twisted and turned, tweaked and

skewed until a favourable result is obtained. It is the responsibility of the data scientist to make sure that the needle she found in the haystack is not just a meaningless bit of straw, and that the scientific process is followed. The processes need to be transparent and open for peer review, testable hypothesis must be created, and all effort must be made to ensure the results are accurate.

3 Introduction to Apache™Hadoop™

In this report the Hadoop framework has been mentioned several times in the context of businesses successfully leveraging big data technology. Apache Hadoop is a project that develops a library of open source software for scalable and reliable distributed computing. It is capable of spreading huge data sets across a cluster of computers that can number in the thousands, and allows access to them using a simple programming model. Each computer provides its own storage and computational ability, and the library takes care of managing them, detecting failures, and making sure the service is highly-available.[18] At the core of Hadoop are the Hadoop Distributed File System (HDFS™) and Hadoop MapReduce, and on top of them various software packages have been developed to fulfil specific needs and requirements ranging from easy access for non-programmers to machine learning solutions.

Businesses working with big data problems are trying to build information platforms that have rich APIs, are designed around enabling easy data exploration, and accept all data formats, no matter how messy or rapidly changing. Hadoop can be viewed as a "one-stop information platform" which enables agile[1] data analysis.[9] While most Hadoop code will be written in Java it does allow for using various other programming languages.

3.1 Short History of Hadoop

Doug Cutting created the text search library Apache Lucene which included the open source web search engine Apache Nutch. Hadoop was created by Cutting as part of the Nutch project and later became a project of its own. The evolution of Hadoop was heavily influenced by the technologies developed and introduced to the world by Google. In 2002 the Nutch project hit the wall of being unable to scale to the billions of pages on the web, but in 2003 Google published a paper on its distributed file system, the GFS, which enabled data storage across distributed machines. The Nutch team then set out to create an open source implementation of this system, the Nutch Distributed File System (NDFS), which today is the Hadoop Distributed File System (HDFS). Google further helped the development of Nutch in 2004 when they published their paper on MapReduce, a distributed platform that runs on top of the distributed file system and enables heavy number-crunching. By 2005 the Nutch developers had created their own implementation and ported all their algorithms to use MapReduce and NDFS. Early the next year this work was taken from Nutch and the independent subproject Apache Hadoop created around it.[19, p. 9-10][2]

Yahoo picked up Cutting and the Hadoop project, kept it open source, and impressed with its simplicity committed to transferring its search infrastructure to Hadoop while foreseeing the future opportunity of this becoming a more general purpose

[1] agile in the same sense as agile software development.

[2] All of the information in this paragraph obtained from Hadoop: The Definitive Guide, as cited at the paragraph end.

technology.[12] In spring 2008 Yahoo held the first Hadoop developer summit where they expected 100 guests but 350 showed up; in 2009 the attendance was double that, and by then Hadoop was being used by big names like Facebook, eBay and even Microsoft.[12]

3.2 Hadoop Distributed File System

A distributed file system manages file storage across a network of machines called nodes. This comes with a set of complex challenges, chief among them being the ability of the system to withstand node failure without losing data. The Hadoop Distributed Filesystem (HDFS™) was developed following Google's publication of a report on its Google File System (GFS)[3]. HDFS "is a file system designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware"[19, p. 41] This means that it can effectively handle files in the terabyte or even petabyte size range where the main processing pattern takes the form of writing the data set once and then reading large portions of it multiple times with minimal read latency. HDFS can use commonly available hardware and is designed to withstand the associated higher probability of hardware failure without interrupting operations through fault detection and automatic recovery. This means that HDFS is not suitable for low-latency data access, many small files, multiple writers, or arbitrary file modifications. The last two relate to how file writers work. Files can be written to by a single writer which always writes at the end of the file, but multiple writers are not supported, and neither is writing at offsets into files.[19, p. 42]

3.2.1 Blocks

Much like file systems and disks work on the level of blocks HDFS also operates on this level, but the blocks are much larger, typically 64 or 128 MB.[19, p. 43] The files are broken down into these blocks which are then stored in the file system, but a file too small to fit in a single block will not take up the full block size. The reasoning behind having such large blocks is to minimize the time cost of file seeks in the sense that if seeking to a block takes 10ms and it is then read at 100 MB/s, the seek time is just 1% of the transfer time.

The main advantages of using blocks to abstract file storage in a distributed file system are the lifting of file size limitation, simpler storage, and replication. A file can be larger than any single disk in the storage system and its blocks can be spread across any number of disks. The storage management can further be simplified through block abstraction and the elimination of metadata which is handled by a separate system. Lastly, blocks are well suited for replication when considering fault tolerance and availability because each block can be replicated to a number of separate nodes. If a block becomes unavailable for any reason, an identical copy can be accessed from a different node and the block further replicated to a different node to maintain the replication factor, all without the user ever knowing that a

[3] Available at http://research.google.com/archive/gfs.html

node failed. If a file is frequently accessed it can have a higher replication factor to spread the read load across the cluster.[19, p. 43-44]

3.2.2 Namenodes and Datanodes

The cluster nodes are split into a namenode master and datanode workers, both of which are invisible to the user. The namenode manages the file system namespace meaning the file system tree as well as the metadata information for all the files and directories in the tree. The information on block locations for each file is stored in memory on the namenode resulting in the namenode being able to serve these locations very fast.[19, p. 44]

The datanodes take care of the heavy work. They store and return the blocks when requested, and regularly report block storage to the namenode.[19, p. 44] The process of reading data from HDFS is explained visually in Figure 1.

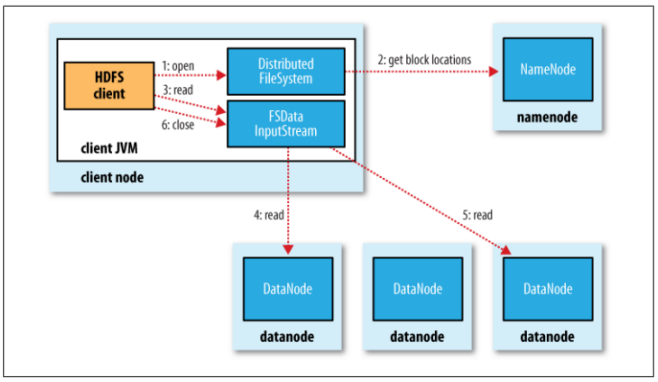

Figure 1: HDFS process for a client reading data from HDFS. The block locations are retrieved from the namenode and a stream is created to the datanodes. Image source: Hadoop: The Definitive Guide, 2nd ed., p.63.

Having just a single namenode means there is a single point of failure in the system. If the namenode fails then the system becomes unusable and the file information is lost. To avoid this kind of catastrophe the system can be configured to store the file system tree and metadata information in multiple locations both locally and on remotely. A secondary namenode can also be run alongside the primary one, periodically synchronizing with the main one and keeping a copy of the namespace data. If the main namenode fails then the secondary one can be synchronized and take over operations, but this process must be performed manually.[19, p. 45][6]

3.3 MapReduce

Google described its original implementation of MapReduce as follows:

> "Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate

> key. [...] Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system."[11]

The MapReduce implementation in Hadoop was inspired by Google's MapReduce publication. Google's motivation was to process many tasks at the same time that would churn through lots of data using hundreds or thousands of CPUs in a simple manner.[11] MapReduce provides: Automatic parallelization and distribution; fault-tolerance; I/O scheduling; and status and monitoring.

A user defines a MapReduce job. It is composed of the input data, the MapReduce program itself, and some configuration settings. Hadoop takes the job, splits it into map tasks and reduce tasks, distributes them evenly across the cluster, and reduces the execution results into a single answer. If the splits are small and many then processing them in parallel takes a short time and the cluster can be balanced well, but the downside is that smaller splits increase the management overhead and increase the split and task creation time.[19, p. 28] A good split size tends to be the size of an HDFS block (typically 64 or 128MB), and Hadoop prefers assigning a map task to nodes where the data resides in what is termed *data locality optimization*.[19, p. 28] The reason for using the size of a HDFS block for data splits is that "it is the largest size of input that can be guaranteed to be stored on a single node."[19, p. 28] When the map tasks finish their results are usually transferred across the network, merged, and fed to the reduce task(s) which then typically store the results in HDFS.

The job is managed by two types of nodes: a single jobtracker and several tasktrackers. The jobtracker coordinates and schedules jobs on the tasktrackers who then report back to the jobtracker on their status and the job status. Fault tolerance is handled by the jobtracker when failure is detected, and the map and reduce tasks on that tasktracker are rescheduled on a different machine.[19, p. 28] As an example of the robustness of this system a MapReduce job at Google ran to completion even though 1600 of the 1800 tasktrackers failed.[11] Figure 2 shows the flow of a MapReduce job with three mappers and two reducers.

MapReduce is suitable for working through large amounts of unstructured data because the programmer decides how the data will we interpreted at runtime and writes the map and reduce functions which define the mapping of key-value pairs to other key-value pairs. These functions can be used unchanged on small and enormous datasets, and the time required increases linearly with the size of the dataset and decreases linearly with the number of machines.[19, p. 6] Hadoop further provides APIs to MapReduce that allow programmers two write their map and reduce functions in languages other than Java. These APIs are called *Hadoop Pipes* if using C++ and *Hadoop Streaming* for other languages. Hadoop Streaming uses Unix standard streams to interface between the program and Hadoop so any programming language that support standard input and output can be used.[19, p. 33,37]

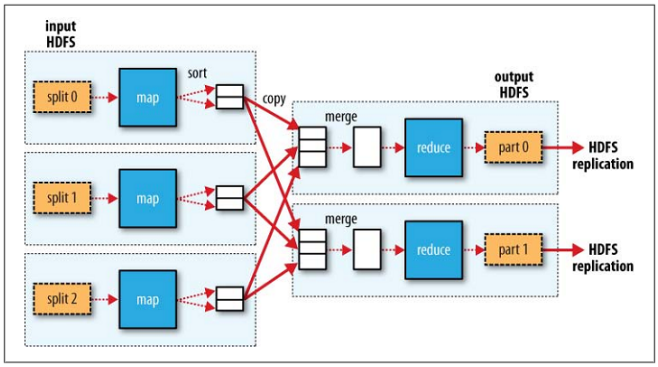

Figure 2: MapReduce data flow with multiple reduce tasks. Image source: Hadoop: The Definitive Guide, 2nd ed., p.30.

3.4 The Hadoop Ecosystem

The two core components of Hadoop are HDFS and MapReduce which collectively provide high-bandwidth storage capable of self-healing, and fault-tolerant distributed computing. This is the essence of what Hadoop provides: "a reliable shared storage and analysis system" with HDFS for storage and MapReduce for analysis.[19, p. 4]

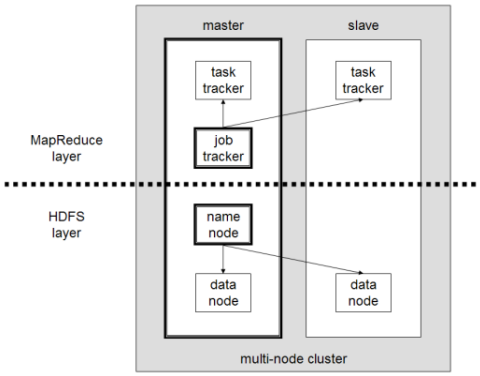

Figure 3: Simplified representation of the MapReduce and HDFS layers on a multi node system. Image source: `http://en.wikipedia.org/wiki/File:Hadoop_1.png`

On top of this there is a whole ecosystem of Hadoop projects created for specific data access or processing needs[19, p. 12-13][18]:

Common
: A set of components and interfaces for distributed file systems and general I/O (serialization, Java RPC, persistent data structures)

Avro
: A serialization system for efficient, cross-language RPC, and persistent data storage.

Cassandra
: A scalable multi-master database with no single point of failure.

Chukwa
: A data collection system for managing large distributed systems.

HBase
: A scalable distributed, column-oriented database. HBase uses HDFS for its underlying storage, and supports both batch-style computations using MapReduce and point queries (random reads).

Hive
: A distributed data warehouse. Hive manages data stored in HDFS and provides a query language based on SQL (translated to MapReduce jobs) for querying the data.

Mahout
: A Scalable machine learning and data mining library.

Pig
: A data flow language and execution environment for exploring very large datasets. Runs on HDFS and MapReduce clusters.

Sqoop
: A tool for efficiently moving data between relational databases and HDFS.

ZooKeeper
: A distributed, highly available coordination service. ZooKeeper provides primitives such as distributed locks that can be used for building distributed applications.

There are many more that are not mentioned in the above list. Hadoop has reached such maturity and become such an industry standard for manipulating huge datasets that if a user has a specific needs it is likely that software that fulfils that need is available for Hadoop.

4 CCP hf

4.1 History of CCP

CCP is an Icelandic game developer and publisher that focuses on Massively Multiplayer Online games (MMOs). CCP was founded in 1997 under the name Loki Margmidlun with the purpose of eventually creating and publishing a space-based MMO titled EVE Online: The Second Genesis. Loki Margmidlun changed its name to CCP in 1999 and released EVE Online in 2003. CCP has grown fast in the years after the release of EVE Online and currently has development offices in Reykjavik, Iceland; Newcastle, UK; Atlanta; USA; and Shanghai, China. CCP currently employs nearly 500 people worldwide who are working on three MMO games within the company. These are EVE Online, Dust 514 and World of Darkness.

4.2 The Games

4.2.1 EVE Online

The MMO game EVE Online was developed by CCP and released in 2003. Whereas most MMOs have multiple servers and separate their players by region and subregion, the actions of the now nearly 400 thousand active EVE Online players have since 2003 been recorded and persisted on a single server[4]. This "one world" or single shard design has always been the core of the EVE online game design. With the addition of DUST 514 the concept of single shard moves more towards the notion of One Universe.

The EVE Online game world consists of over 7,500 star systems, but is essentially split into three parts: (1) high security solar systems that players cannot claim ownership of and provide relative safety from player aggression; (2) low security solar systems where players cannot claim ownership and provide very limited safety from player aggression; and (3) zero security solar systems that players can claim ownership of through war and conquest against other players. A player of EVE Online can pursue a career in economy, industry, warfare, science, technology, or whatever blend she chooses. The game is presented as a sandbox where the game provides the tools and the players create the experiences. Players can band together to form corporations which can then join forces under an alliance. Corporations can declare war on each other and can claim territory in zero security space in the name of their alliance, but players and corporations can just as well operate in the safety of high security space if they so choose. The main draws of the game are the dynamic and entirely player-driven economy of the game and the emergent gameplay created by huge groups of players where the actions of one group, or even one player, can echo across the universe. The game's current monetization strategy is traditional in the sense that players can purchase subscriptions of a multiple of 30 days directly from CCP and retailers.

[4]EVE Online in China is, however, hosted on a separate server in accordance with Chinese law.

4.2.2 Dust 514

Dust 514 is currently in development at CCP's Shanghai and Newcastle offices. Dust 514 will be a high quality free-to-play MMO First Person Shooter for the SONY PlayStation 3 gaming console and is set for release in 2012. Dust 514 will be an integral part of the EVE Online universe and the battles fought there will take place on planets in EVE Online. This is in fact how corporations in EVE Online will be able to take ownership of planets—through wars and battles fought in Dust 514. The actions of players in Dust 514 can thus directly impact players in EVE Online, and vice versa. The games will also share some parts of the in-game market where items and currency can flow between the two games, but the design of that aspect has not been finalized at this time.

The monetization strategy in Dust 514 will centre around a free-to-play model where players do not need to purchase the game to participate. The revenue comes from players purchasing in-game currency which the players can then use to purchase optional things for their characters such as better equipment and weapons. This model is still under construction and the team is actively evolving the monetization strategy.

Dust 514 presents numerous challenges to CCP Games. The communication between EVE and Dust 514 is a world-first for combining a PC MMO with a console based MMO. The servers where the battles in Dust take place, the battleservers, will be distributed all over the world. After a battle, certain information will be transferred from a battleserver to the EVE Online server, but most of the really important behavioural data will remain on the battleservers. Managing, centralizing, and analysing all that data will be a major challenge, especially considering that the PlayStation 3 console has sold over 60 million copies [17] putting the potential user base of Dust 514 in the tens of millions. The amount of data generated in that best-case scenario would soon dwarf the amount of data collected in EVE Online.

4.2.3 World of Darkness

CCP and White Wolf Publishing merged in 2006. White Wolf has published several role-playing games all of which take place in the fictional World of Darkness universe, and CCP is now developing an MMO based on that universe titled World of Darkness. The game is primarily being developed in the Atlanta studio and is too far from release to be included in this report.

4.3 Current Data Analysis Efforts

In 2007 there was no meaningful way for CCP to get data from or perform analysis on the EVE Online server except during lunchtime, right after EVE's scheduled daily downtime. Today a fresh backup of EVE Online is created daily at midnight and used for research, analysis, and investigations. This research database belongs to the Research & Statistics department which is also the main user, but several others have access to it as well.

Because of the performance requirements associated with operating EVE Online on a single SQL Server, the logs have been carefully crafted to be as slim as possible. It has tended to be the Customer Service and Research & Statistics departments pushing for increased logging for better analysis and customer service, with the database administrators and programmers pushing back in favour of cluster performance and stability.

Various departments in the Reykjavik office have their own measurements, dashboards or metrics that they actively monitor when it comes to their own performance or various aspects of EVE Online. These are scattered all around and do not form a coherent whole and are therefore not covered in this chapter.

4.3.1 Resarch & Statistics

The Research & Statistics (R&S) department has the objective of serving CCP and the subscribers with accurate and actionable information on player behaviour in the virtual societies operated by CCP. The main objective of R&S is to provide information to players of EVE Online as well as to CCP as an operator of EVE Online to cope with the increased complexity of the EVE Universe. For the players, the focus is on providing economic and socio-economic data and analysis in order to enhance their understanding of the economic system within EVE Online. For CCP, the focus is on providing actionable information on behaviour within EVE Online in order to facilitate its operation as well as assisting in future development of its other games. R&S has published results from data analysis and socio-economic research online as well as in printed newsletters such as the quarterly economic report. The department also participates in commercial events and prepares material for any kind of event if needed. R&S also handles all communication with academic institutions interested in EVE Online related research projects.

R&S is split into a data warehouse team, game behaviour team, subscriber team, and a survey team. The data warehouse team maintains a data warehouse that integrates various data sources, provides cleaned historical data for ad hoc reporting and analysis, and builds OLAP cubes for data analysis. The available OLAP cubes provide flexible analytical possibilities for a range of subject areas such as subscription sales, game feature usage, combat, production, players, and customer support.

The game behaviour team focuses on socio-economic research within EVE Online. The team monitors and reports on the EVE Online market, production, new player activity, and combat activity. This team also closely monitors the PLEX[5] market and assists with PLEX sales offers. The team also provides metrics and analysis for other EVE development teams by request.

[5] A 30 Day Pilot Lisene Extension (PLEX) is a tradable in-game item that can be used to add 30 days to a player's subscription, or sold for a hefty sum on the in-game market. Additionally, the player can give the PLEX back to CCP in exchange for special currency that can be used to purchase clothing for his or her avatar in the game. This enables some players to essentially play for free by purchasing PLEX on the in-game market, but each PLEX is originally injected into the game by another player purchasing it directly from CCP.

The subscriber team focuses on churn analysis, reactivations and retention efforts for CCP's customers. The churn analysis is the basis for the subscriber forecast and also provides information that can be used to optimize reactivation and retention campaigns.

The survey team conducts various surveys for CCP. All of these surveys provide valuable information on how the customers perceive EVE Online at any given time and what might be immediate problems that have to be fixed. The survey team can also provide other EVE teams with custom made surveys that tackle issues directly relevant to that specific development team

4.3.2 Key Performance Indicators

CCP and EVE Online have several well established Key Performance Indicators (KPIs). These cover the EVE Online economy, players, trials, Customer Support, database operations, the financial side of things, etc. These are updated monthly and collected in a dashboard of sorts where they give a sense of the general performance and health of the EVE Online operation over time. These metrics have evolved over the years and in many cases there is significant analytical work behind them.

4.3.3 EVE Metrics

EVE Metrics sprung out of team formed to focus on metrics and to prototype alternative logging with the aim of taking huge amounts of otherwise discarded data generated in EVE Online and turning it into information to help EVE Online development in becoming more metrics-driven. Their work resulted in EVE Metrics, which is a framework that allows individuals or teams to collect their own basic metrics originating from the EVE Online research server and the EVE Metrics database and save them as 'counters'. These counters are populated daily and made visually accessible as reports on an internal website where users can manipulate the data in simple ways. Users can then create named collections of reports to present single or multiple counters in various fashions. As an example, this allows for the creation of a collection showing counters relating to daily player participation in various gameplay elements, or the display of all sorts of activity in a single system. The intention behind this project is to awaken people to start using the data to help in their decision making and feature monitoring. The counter creation is easy for programmers and any user can create a collection.

At the core of this project are numerous data hooks placed in the server code to log events in an unstructured fashion to text files in far more quantity than previously possible while causing a tolerable hit on the server code. This project has succeeded in making a cheap logging solution that would otherwise have been very expensive and also represents CCP's first steps towards big data solutions. EVE Metrics has been successful in getting more people interested in the data. People like the simplicity of the system and how it allows them monitor all sorts of things and to build the data hooks directly into their new features.

5 A Survey of CCP's Big Data Requirements

CCP has decided to adopt big data technology and analysis into the way they operate and maintain their games. This comes not only as a reaction to the data needs presented by Dust 514's business strategy, but also as a reaction to the company's strengthened emphasis on being data-driven. Many within the company have an opinion on why CCP is going this route, but their reasons, hopes and expectations have not been consolidated and discussed. This chapter is intended to be such a consolidation.

5.1 Interview Methodology

For this project the decision was made to interview individuals from three groups within the company: management, influencers, and likely users of big data technology. These groups are referred to in the following text as managers, influencers and users. Each group included people from both the EVE Online and Dust 514 sides of the company. This selection was chosen in order to get strategic as well as practical perspectives on why CCP wants to use big data technology and what challenges it is intended to help solve. The interviewees are listed in Table 2 and a more detailed description of each of them is presented in Appendix A.

The interview questions are shown in Figure 4. They were chosen and ordered with the intention of creating a flow of people thinking about data related challenges facing CCP today, whether big data technology can help CCP overcome those problems, what opportunities this technology could create, how important this is for CCP, and finally who should be the main users of the technology and information coming out of this type of analysis. Keeping in mind that people have various backgrounds, expertise, and technical know-how, the decision was made to avoid direct technological questions and instead focus on what people would like to get out of the technology.

The interview approach was chosen over sending people a questionnaire because the topic was quite open and the interviewees were very busy. Sending a questionnaire would probably have resulted in a low response rate and annoyance on both sides. An interview also makes it possible for the interviewer to ask for further details and more specifics, which is important when the topic is open and the interviewees have differing opinions or ideas.

All the interviews were recorded and then transcribed. While being a very time consuming process this provided a clear overview of each interview and allowed for much easier extraction of information and comparison with other interviewees.

5.2 Interview Results

The interviews had often overlapping answers to some of these questions, but in many cases they went on different tangents that reflected their individual experience and specialization. The variety of answers is consolidated here below in a few

Role	Name	Title	Game
Management	Hilmar Veigar Pétursson	Chief Executive Officer	N/A
	Jón Hörðdal Jónasson	Chief Operations Officer and CCP Asia Manager	Dust 514
	Halldór Fannar	Chief Technology Officer	N/A
	Jon Lander	Senior Producer	EVE Online
	Brandon Laurino	Executive Producer	Dust 514
Influencers	Andie Nordgren	Technical Producer	EVE Online
	Eyjólfur Guðmundsson and Lead Economist	Director of Research & Statistics	EVE Online
	Jón Bjarnason	Programmer	EVE Online
Users	Ingólfur V. Ævarsson	Director of Customer Lifecycle Management	EVE Online
	Kristoffer Touborg	Lead Game Designer	EVE Online
	Eino Joas	Game Designer and Economist	Dust 514
	Ben Cockerill	Producer	Dust 514

Table 2: List of interviewees at CCP with their roles, titles, and game.

categories. These are the data gathering and utilization, understanding of big data, why CCP is going with big data, how important this is perceived, the associated challenges, and finally who should be the users of the information coming out of big data analysis. The interviewees are referred to by their role (manager, influencer, user) and efforts made to not directly identify who is being quoted. In some cases an individual will be identified in terms of his role and game he is working on, but that is done to emphasise where the opinions differ between the two projects.

5.2.1 Current Data Gathering and Utilization Challenges

Having a single server in EVE Online simplifies a lot of the analysis efforts compared to what others in the industry have to deal with. However, a recurring theme from all interviewed regarding EVE Online was that having a single SQL Server at the core of the game creates severe obstacles for the data creation and analysis efforts. The main obstacle, until the previously mentioned EVE Metrics project[6] alleviated some of it, has been logging enough behavioural data for effective analysis. They all said that from a gameplay performance and hard disk cost perspective it has simply not been feasible to store as much data as many wanted.

The nature of SQL Server requires the data inserted to be mostly structured which creates a second obstacle. One of the managers described it as follows. For every aspect of the game the decision must be made whether or not it should be logged, and if it is logged then how the data should be structured in the relational database. This process requires the people who may later need to analyse the data to "define their analytical needs and priorities in advance to the people who implement the structures in the database." Later on there may emerge a data need they had not anticipated, resulting in the available data not matching the actual need. The challenge is therefore "to make the right kind of data available and to empower the people

[6]see Section 4.3.3 on EVE Metrics

1. What is your role/title within CCP?
2. What type of problems or challenges is CCP faced with today in terms of utilizing information / data obtained via our current technologies?
3. What is your understanding of Big Data / Hadoop and the type of problems it can solve?
4. Do you believe CCP should adopt Big Data / Hadoop solutions?
   (a) Do you believe Big Data / Hadoop can help solve the problems you previously mentioned?
5. How would you prioritize Big Data / Hadoop amongst the projects that are under your control?
   (a) Perhaps on a scale of 1 -10 ?
   (b) What projects are more important?
   (c) Will Big Data / Hadoop have any impact on these projects?
6. Who do you think should be main users of the information / actions coming out of big data / Hadoop analytics?
7. Any other comments or issues you believe should be addressed in the implementation of a big data / Hadoop solution?

Figure 4: Interview Questions

who need it to be able to access it." They should not first have to wait for insufficient structures to be modified and then for the new logs to be written because by then the moment may have passed and the initial question been rendered irrelevant. An influencer said that although logging represents around 75% of the inserts to the SQL Server today, this technology is not very well suited for logging because of the database's rigid structural requirements and performance considerations. The rigid data structure was thought up in a different era, and "while the data is interesting it does not match what CCP needs right now." Specific cases can be solved using current technology to satisfy logging and analysis needs, but in their opinion CCP needs a setup where this is "so cheap in terms of logging, computations, and storage that they can start doing it for as much of the game as possible."

The concern for some of the users is not only about logging the right data or having the right technology in place. In their opinion CCP has simply not been able to dedicate enough resources on performing many advanced analyses on the already available data. The EVE SQL Server and the server archives have a lot of valuable data that they said has been underutilized. For the people tasked with crunching through the data the first challenge they face is that there is so much data available that transforming it into actionable information becomes the main problem. The second challenge is the time constraints which lead to "picking and choosing what problem we can implement and actually push through the company." Gathering the right data can take a lot of time and the complexity of the database means that there are not a lot of people capable of finding it and using it in a meaningful way. Having the data easily accessible is therefore identified as one of the key challenges, along with maturing the processes and pipelines behind the analytical work.

Challenge	Effects
Single SQL Server	Minimal Logging Strictly structured data Logs do not match analytical needs
Insufficient analytical resources	Current efforts only cover a fraction of the available data Limits the scope of projects Analytical pipelines underdeveloped
Teams not data-motivated	Difficult to get them interested in data Few know how to use the tools
Difficult to find the data	Very few can navigate the database People give up
Dust 514 requires more data storage and a more rapid analysis loop	Analytical solutions from EVE cannot scale Current storage solution insufficient Need up-to-the-minute information

Table 3: Data Collection and Utilization Challenges at CCP

Some said that in past years decision making at CCP was often based on intuition or experience, with lesser emphasis placed on gathering data to aid with the decision-making. They added that this has changed over the years, but development teams have not yet been properly motivated to create their own trends and analysis of their game features. The challenge is "getting people interested in the data and empowering them" with the abilities to implement data hooks and perform analysis on their features after they are added to the game. People are not accustomed to the fact that they can do this. But getting people interested is just one side of the problem. Once they get interested, the problems facing most of them revolve around usability and how the data is presented. In many cases they lack the technical or mathematical knowledge to trudge through the data themselves or effectively use tools that provide a window on the raw data. For them it is most important to "have the information presented in a way that is easy for them to use." They expressed the need to be served the information either through reports or through tools that make it intuitive and easy for them to extract information. For them the main challenge is usability improvements, being able to get more data and information faster, and being able to customize how they view it.

With nine years of EVE Online as a subscription based game, CCP has not had a desperate need for deep analysis or up-to-the-minute data, but this will change dramatically with Dust 514. Interviewees on that side said the game's free-to-play monetization strategy requires deeper analysis than previously performed, where for example the battle behaviour of players, their interactions with the UI, and their purchase behaviour needs to be analysed in order to improve and fine tune the game experience, increase player retention[7], and increase revenue. The user base is predicted to be many times that of EVE Online, player activity will be more rapid,

[7]Retention is the proportion of players who keep playing the game versus those who quit.

and the data will no longer be all in one place. They know that the analytical solutions available from the EVE Online side will not scale to fit all analytical needs in Dust, and that the sheer amount of generated data will not fit in the current storage solution.

Some managers said that with the upcoming big data integration the problem will shift from not having detailed enough data to work with, to one of ensuring that the data is clean, consistent, and making sure that people understand it well enough to be able to make decisions based on it.

A summary of the data collection and utilization challenges at CCP is presented in Table 3.

5.2.2 Understanding of Big Data and the Technology

"Big data" is a phrase that people sometimes thrown around as some sort of end-all solution—that they will be able to "press the 'big data' button" and solve everything. There was some hint of that sentiment during the interview process with the sheer scale of problems and opportunities put forward that big data technology could help solve. But for the most part the consensus was that big data allows people to: store orders of magnitude more data than before, crunch it in a cost-effective manner, and make it more accessible for people and tools. Everyone believed CCP should implement big data solutions and make serious efforts to adopt the analysis into the Build-Measure-Learn process of development. No one believed big data would bring about some sort of paradigm shift or revolutionise the way CCP operates, but viewed it more as a specialized tool in their ever expanding problem-solving toolbox. However, the incredible effort and expertise required to achieve some of the lofty goals put forward was not really touched upon.

5.2.3 Why is CCP Going the Big Data Route?

Everyone interviewed believed CCP should adopt big data in a way that makes sense for the company. Big data would not radically change the way CCP operates EVE Online, it would simply increase what they can do with data, fill in important information gaps, and provide opportunities for new features and services. EVE has been operated for 9 years without big data technology so getting it for EVE Online is not viewed by all as critical right now. It would help, but there is not a pressing need. Dust 514 is an entirely different matter because there they need to continually manage and fine tune all levels of the game experience to make the business plan work. According to the interviewees on the Dust side big data capabilities may not be absolutely critical before the game launches later this year, but it will soon afterwards become a "must have". This means they must begin equipping themselves for big data analysis very soon.

Solve current problems
Big data technology was viewed by most on both the EVE Online and Dust 514 sides as not only a strong analytical tool, but also as a tool that could enable more effective

information delivery to the people who need it. It would enable unstructured logging where the usefulness of the data is perhaps not yet known, but when questions arise the data could be structured and queried to answer them. Sometimes the EVE Online players tell the developers that the game has a specific problem or the developers may suspect a problem, and with this type of logging and analytical capabilities it could be investigated in more detail than is currently possible. They could "see better where things are not streamlined, gauge participation, and see how certain aspects of the game are used", instead of implementing a solution based on what they have heard or the limited data currently available. It would allow them to have access to as much information as possible when they make their decisions, and see in more detail the effects of those decisions.

A manager said that CCP has been losing out on the opportunity of tracking a lot of very interesting data in EVE Online. An example he mentioned was that they are continually modifying the game by adding, removing, and changing things, and it would be "very beneficial for us if we could have up to date metrics on what those effects are." The only way for them to track this is to log out much more data than they have done, and then big data technology could help them crunch it and provide those metrics. The EVE Metrics initiative is already providing some of this, but there the data cannot be transformed into all the different kinds of information the users need. Another example mentioned was that Customer Service (CS) representatives are constantly looking at logs in the database to investigate cases brought up by the players themselves or flagged by internal systems. This work can be very complicated and time consuming where a single investigation can sometimes extend for an hour or more. The EVE Metrics project has already filled some of the logging holes that CS previously had to work around, but "giving them the additional ability to transform and pivot the data around the things they find interesting" has the potential to hasten investigations and in general make the work of CS much easier. This would in their opinion only be possible with big data technology. CCP is acquiring a lot more data than before, they just lack the infrastructure to manipulate it.

When it comes to the lean build-measure-learn loop, a manager said that CCP does a lot of building but they are not doing that much measuring, partly because the current infrastructure does not support it well. This means that the third part, learn, is in his opinion kind of hurting. EVE Metrics is helping them greatly to complete the loop, and the belief is that with big data solutions they can take it even further by enabling easier access to data for people and tools. Many decisions up until now have been educated guesswork so people are looking forward to having more data to back up those decisions.

Dust 514 is being designed in a different era than EVE Online. Its operational requirements will make it possible to capture a lot of data from the outset, which sets it up for big data analysis. There will be a huge amount of behavioural data distributed across an undefined scale of computers throughout the world, and "being able to do real-time and behavioural analytics on this data in a cost-effective manner is at the core of the challenges" facing the Dust 514 developers. Very basic examples mentioned were analysing battle results and user interface (UI) usage. Basic battle analysis is currently possible, but they need to perform much more detailed analysis,

Current problems big data could solve
Unstructured logging
2+ orders of magnitude more logging
More detailed analyses of usage and performance
Better information delivery
Better information access for decision making
Up to date metrics
Enables better Customer Support
Helps complete the build-measure-learn loop
Supports the scale, storage, cost, and analytical requirements of Dust 514

Table 4: List of current problems CCP could solve with big data technology

for example where players were when they achieved the most kills, what was the rate of bullets fired in different areas by players in different alliances, and how did people move around the map. As for the UI, from a game design and UI design point of view it is for example very important for them to be able to see "which parts of the UI are most/least used and how problems are linked to people quitting." As part of the monetization strategy they want to measure things such as "how much time players spend in the market, purchasing behaviour, consumption behaviour, what items are popular/unpopular and why, seeing the results of introducing new items", and so on. The logged amount of data needed is of such a scale that they could never really aggregate and crunch it on a single machine. The interviewers said this kind of analysis was critical for much of the business goals required to manage a free to play game as well as needed. The fine tuning of Dust 514 and the fine tuning of the business model depends on their analytical capabilities. For games like EVE Online that data was said to be very valuable because it allows them to optimise their operations much more than they otherwise could, but they said that for a game like Dust 514 it is absolutely critical, and the Dust developers believe they will only be able to effectively analyse the data once they get a tool such as Hadoop.

A summary of the current challenges that big data could help CCP solve is presented in Table 4.

New Opportunities

In extreme cases of information need, the live EVE Online server can be queried for data, but that is avoided at all cost as it may impact gameplay. A manager said that here has not been much use in having data less than a day old to work with. However, the game world is a constantly shifting environment where players are interacting with the game and each other 23 hours of each day. The question he posed was "how can CCP use the knowledge of what goes on in the game world to make the product better and make features that today would be too expensive or complicated?" In his opinion there are opportunities to make the virtual world seem more dynamic and therefore real by using data in an automated way. As an example, every hour players kill thousands of computer controlled bad guys that

are supposed to be parts of big organized groups. Despite thousands of casualties they keep reappearing in the same locations. They could react to the slaughter by relocating themselves and thus making the world feel more realistic to the players. It is too expensive to use the current technology to crunch data and immediately use the results to alter the game, but he feels that big data technology opens up for this kind of feedback loop where the actions of the players could have a near-immediate impact on the game systems.

He went on to say that the feedback loop should not be limited to only modifying how the game systems are functioning. CCP's statement and company vision is: "making virtual worlds more meaningful than real life", and all of CCP's games are about players creating these meaningful social networks and experiences. Many mentioned that most EVE Online players play the game because they are interested in the social aspect of the game. One interviewee likened this to "people's actions making footprints in the sand, then someone else comes across those footprints and feels that this world is alive." They want to make those footprints as big as possible so that "many people see them and realize that this is a living breathing world." Examples mentioned by the group were showing each player a personalised feed of things that are happening in the universe relevant to that player, showing extended lifetime statistics of the player's interactions in the world, showing the complete history of the player's spaceship, giving recommendations for skill training, and suggesting gameplay the player might be interested in trying. If a player logs in to the game it would be great to have a ticker, similar to what Facebook has, giving some information on "what his friends are doing or what his five thousand alliance members are doing." If they can make this much more personal and real-time then it could have a powerful effect on players. Stepping away from the single player, the main content of EVE is what other players do, and if they can channel this back to the players and make interesting things out of it then the players could generate even more content. They could generate a "Facebook/Twitter effect of this velocity of information and interaction flowing much more rapidly within EVE." These are things that are hard for CCP to do today in EVE Online because they are expensive and not really designed for, so work is required on that end, but Dust 514's design opens up for more intuitively changing the game world in real-time and serving data back to the players.

EVE Online and Dust 514 both take place in the EVE universe, and even though they are on different platforms (EVE Online on PC and Dust 514 on PlayStation 3) the hope is that Dust players will try out EVE Online and vice versa. The interviewees said that having the ability to not only present players with live information on what is going on in their game but also across the games could represent a way to get players interested in the other game. Even if an EVE Online player is not interested in Dust 514 he might see a line in his information ticker saying that district X on planet Y just fell to Amarr forces, and he might think to himself "this is of no interest to me", but yet he would realise that a lot is going on in this world.

In Dust 514 the question arose whether big data analysis could be used as part of the monetization strategy in a more direct way. Perhaps alliances or corporations would purchase battle behavioural data such as heat-maps or item consumption on their own members, or even their opponents. EVE Online players might want something

New Opportunities
Automated systems that make the world seem more dynamic Enhance the social experience and give feeling of living breating world Cross live information between EVE Online and Dust 514 Serve more data back to players, they can create incredible things New monetization opportunities More and new types of analyses: - Gameplay behavioural analyses - Understand player types, the paths they take and why - Understand what leads a player to a purchase decision - More effective reactivation campaigns Understand players better

Table 5: Shows new opportunities for CCP using big data technology

similar on their end. The opinion expressed on the Dust side was that there are countless opportunities for utilizing big data analysis to both create value for the players and revenue streams for CCP, they just needed to find the right ones.

Those who are currently creating analyses and crunching data on player behaviour mentioned many needs that big data technology could help fulfil. They said CCP needs to get more information from the games on the behaviour of their players. They need to be able to understand social networking and behaviour in much more detail, both within the games and around them. They need to understand "how people are communicating, even outside of the client, and identify who the 'connectors' are" that drive their fellow players. They need to understand "what makes people tick once they make a purchase decision, understand the path that people take from the time they see the game or they hear about it until they become a long term subscriber" or a purchaser. They need to explain "why a particular person or group of people were not likely to be long term subscribers and hence losing them was just fine, and also need to know why didn't get a hold of the people we do know should be interested in the game." They need to be able to cluster the players better than they can today to see what kind of players they are to help with marketing campaigns where they can create custom promotions and identify who to contact. In the end, one influencer said, CCP "needs to understand people better and at a much faster pace than we have been able to do until now." CCP is just beginning to scratch the surface of how big data can help them and the expectation is that in a years' time from now they will be using it for things they haven't even imagined today.

A summary of the new opportunities big data presents for CCP is presented in Table 5.

5.2.4 How Important is Big Data Analysis for CCP?

CCP has spent over 10 years on making EVE Online align with their vision, but limited resources on their analytical pipelines and processes. When thinking of the importance of big data for CCP one has to think of it in terms of both Dust 514 and EVE Online. Dust 514 is being designed with detailed analytical possibilities in mind and it is the belief of all interviewed that Dust 514 has a clear and present need for big data analysis. While EVE Online does not have a pressing need right now for big data solutions, and is not being built with information extraction as a core foundation yet, the managers said they could imagine that in 18 months it may well have become absolutely essential. The opinion nearly across the board was that big data technology would be very beneficial for EVE Online, but critically important for Dust. One of them said that using big data is part of a strategy that competing MMO game developers are already making use of and if CCP does not do it then it would "fall behind [...] there is just no question about it." One of the influencers added that they have to keep up with the present priorities but at the same time equip themselves with new capabilities so that in the future they could be in a position where it is cheap for them to use big data technology. Otherwise they would "keep being dragged down and constricted in the type of solutions we can come up with by the constraints of the existing technologies." This would pose an unacceptable strategic risk because it would restrict them, development wise, to the current technological range, which would cause them to "lag further and further behind the competition", and in the worst case scenario they would discover that too late.

When asked to put things into perspective and prioritize, it emerged that right now several things for EVE Online and Dust 514 are considered more important than getting big data up and running. For Dust 514, which is currently several months from release, simply getting the hosting infrastructure, hiring the right people, getting processes defined, getting all the game mechanics working, and getting the game to look and play great were mentioned as things that are more important right now because without them there would be no game. One of the influencers said that big data was currently "high on the agenda and needs to be understood, but it's moving up the ladder quite rapidly." By the end of the year he said they will need to have a solution in place for Dust 514 where they can crunch through their data in a timely manner and report it to both managers and customers, so by then the priority would be a 10 out of 10. A manager said that the type of results coming out of big data analysis in Dust 514 could be very valuable for the players and therefore he believed it would become a critical part of the game's monetization strategy.

With EVE Online a big expansion is released every 6 months with smaller releases in between. Many on the EVE Online side said it is currently more important to finish features, get the expansions into the hands of the players and enable more productive development than equipping for analytics. A developer said that of the four tiers of priority it is currently sitting in the third one, but he added that this was due to its fractured nature and did not reflect its actual importance. In their opinion EVE Online does not necessarily need big data solutions to fulfil its business plan this year, but as mentioned above they feel it will become strategically very important for EVE Online in the not so distant future. However a developer

How important is big data for CCP?
Critically important for Dust 514 post release
Beneficial for EVE Online now, important later
Strategic risk of fall behind the competition if CCP equip itself for big data
Sticking with current technologies would limit the types of solutions created. Again a strategic risk.
Need to better measure how the game is used in ways that only big data analysis can solve

Table 6: How important is big data for CCP?

behind EVE Metrics believes that big data is an extremely important project right now because it is necessary to allow CCP to make better decisions in game design, marketing, and management "based on real world data instead of guesses."

When it comes to development the users on both sides said that ultimately the most important thing is "shipping something customer facing", and when it comes to big data it would be used to measure and enhance the games. One of them on the EVE Online side said he did not think CCP should be making decisions based on data but rather use it "to make sure we have all the information available when we make the decision." They would make a feature, and making sure it is ripe would hopefully come from something like big data giving them a complete picture of it. A developer on the Dust 514 side said that the benefit of big data would come sometime after the core features were ready. Then they would start benefiting from it when it comes to optimizing the core. The importance of big data could thus be summed up as "nice to have" for EVE Online right now but moving to a "must have" in the next 12 to 18 months, and a "really want to have" for Dust 514 but moving to a "critical to have" after release.

The interviewees were finally asked to prioritize equipping CCP for tackling big data problems on a scale of 1-10, with 1 being top priority, amongst the projects under their control. The average turned out to be 3.5, but there was a clear difference between the groups. The average for the managers was 3.0, for the influencers it was 3.6, and for the users it was 4.3.

A summary of the importance factors for big data in CCP is presented in Table 6.

5.2.5 The Challenges of Adopting Big Data in CCP

One interviewee talked about how in 2007 there was no nightly EVE Online server backup to do research on so if a lot of data was needed then the server would be queried at noon during the daily server downtime. He said a lot has changed since then—CCP is now updating its data every 24 hours, the company has become quite a bit more data driven, even though it could do much better, and it is now going

to start working on big data solutions for further data operations. He summed it up by saying: "I would say we are on the right path and maybe we just need to speed up a little bit to make available to everyone the information they need." However, members of the users group exposed a problem with this sentiment in that CCP's processes for analysing data are in many ways still quite in their infancy, even though they have improved considerably over the last few years. They said CCP hasn't gotten around to formalizing their analytical pipelines, and this opinion was echoed across all the groups. They said it is one thing to deal with the amount of data being generated, but the other side of the coin is "how do I then report on it, what do I want to report, and how do I close the loop so data coming in is turned into information, how do I analyse and synthesize that information into meaningful actions?" This is a problem the people currently involved in the analytical work fully realize big data technology cannot help them with. They see it can provide them with data and the ability to interact with it, but it won't magically make new analyses available for everyone.

Another challenge that some of the managers mentioned is that with something like Hadoop it is already known that the technology is solid and can enable fantastic work, but adopting a technological solution in a company fails if it is just being implemented because they think this is what they need. Everyone at CCP understands the company needs big data solutions, but some are worried the actual questions that will be answered or created using that technology have not been defined and they therefore cannot say with certainty what technology or approach would be most beneficial. The discussion has been that they know there will be a lot of data, it will be spread out, they should find the "best possible big data solution", put it in place, and then start asking the questions. This approach may work perfectly well, but some do not see it as the ideal starting point. They want to define the breadth and depth of the kind of questions that will be asked and then investigate what kind of big data solution they need to get the answers. If the technology is put in place before the questions have been defined then they might realize that the problems could have been simplified or segmented in a way where this type of technology was not needed. Then they would end up with an over-engineered solution.

A deciding factor is the resources that will be needed in a project such as this. One from the users group on the EVE Online side said that if the investment in big data could pay off in six months then it should absolutely be done, but if it was a bigger project that would take a year or two before benefits would appear then he wanted to postpone it "and use the resources to work on the problems that are currently waiting and we know we can do." Those problems are many and the resource battle is already so hard that they have to prioritize and very carefully pick the ones they can solve. His worry is therefore that big data technology could lead CCP too much into finding projects while spending too little time actually implementing them. Having too many possible projects may be considered a luxury, but the problem even now is that the available projects are so many and only a fraction of them gets implemented. He would therefore much rather have a tool that helps him solve these problems rather than finding new problems to solve, but at the same time he acknowledges the need to do both. Another user on the Dust 514 side echoed this sentiment when he said that "just getting the data, storing it, and making it accessible isn't going to be enough" if the processes and resources are not in place to

get results out of the data. Examples of the new possibilities big data could create for CCP were recommendation features for players based on other players' actions, live data feeds, and using player behaviour to immediately affect game systems. Management acknowledged that creating these solutions and putting them in place would be nothing short of a massive undertaking for the company.

One of the users also advised against going blindly into building this technology and would rather that CCP worked up a full understanding of what kind of resources it takes to implement this, the steps needed for implementation, who is going to use this and how, and whether the people who are going to use this have the time, resources, knowledge and accessibility to use it. In his opinion this is not a "should we do it or not" situation, this is a situation that needs to be carefully thought through before getting into it. He added that if CCP decides to adopt big data technology then it becomes an independent project that they have to spend a lot of resources on over a long period of time. There are a few such in-house projects in CCP, each of which requires several people to work full time on maintenance and continuous development. He said these projects take resources away from CCP's core competencies of "creating cool games that are immersive and fun." Having these in-house projects has been cost effective and beneficial for CCP until now, but he added that when a large chunk of the company is required to maintain them rather than working on the core competencies there comes a point when they must think about whether they should outsource, whether they can get added revenue from these projects (e.g. opening up their billing system for other companies to use), and how the status and resource use will be after 2, 5, or 10 years.

One from the influencers group had an altogether different opinion on the topic of clearly defining projects and goals beforehand. His feeling was that once they get a Hadoop cluster running they should simply start playing around with it and that way discover how they can use it rather than "starting with a large problem which was not solvable before and requires an extensive requirements analysis and a 500 page report." He favoured much more the grass-roots approach and not attacking a broad problem at the outset. In his mind this would simply represent a research spike where a few people "play around and create cool things." After that they could move on to looking at more ambitious projects.

One of the more interesting challenges for CCP is getting people to access and use more data. The EVE Online developers have not necessarily been using data at the heart of their decision making through the years. Data may not be explicitly necessary for all decisions, but CCP's goal now is to focus more on data driven development, data driven decision making, and actionable metrics. The expressed belief is that the people making Dust 514 have a better understanding of the importance of data both because of the business model and because some of them come from a background where they were trying to do more social networking and behavioural analysis. The feeling among some of the managers is that while there are some people who have a clear understanding of the importance of data on the EVE Online side, the majority have not really been exposed to what data analysis could do for them. In spite of huge efforts to convince game design, marketing, and others of the importance of using data to a greater extent it has only had a limited success. One manager said that his experience showed that it is difficult to change

people's behaviour if it's not critical for them, and therefore he was a believer in showing a section of the people the value through different measures and then having it gradually come in place for the others. Part of the plan is therefore to get the data analysis and usage right from day one in Dust 514 and then use those successes to influence people on the EVE Online and World of Darkness sides.

Increasing people's usage of data poses the twofold challenge of getting people interested in the data and empowering them to get access to it. EVE Metrics represents huge steps forward for both of these as the development teams have embraced it and are putting in many data-hooks themselves. One of the developers said he believes this is "something they should do for as much of their work as possible" because of how it allows them to see in much more detail than before how their features are being used and what effects changes have on usage. However, one of the developers of EVE Metrics said that while he was trying very hard to get people to think about data it was not going well enough, with people giving excuses such as having to put this work into sprint planning. The people behind EVE Metrics are not digging for data, EVE Metrics is built as a service that enables the teams themselves to fish for information. This is good for those who are data driven, but within the company there are many who are not or simply lack the time to invest in it.

The current analytical solutions are only covering a fraction of the available data and not many people have the skills to dig beyond them into SQL Server. Even the currently available tools such as accessing OLAP cubes through Excel and using Targit[8] to create reports, analyses and dashboards on top of OLAP cubes were said to strike many potential users as too complicated and time consuming to learn, while EVE Metrics seems to hit somewhat of a sweet spot in terms of ease of use and understanding. However the users also want to be able to slice the data and dig into it a bit more than EVE Metrics allows. On the one hand these users want to be considered as consumers of this data and as such would like to have the analyses served to them in the form of reports and dashboards, but on the other hand they want tools that allow them to "get the information really quickly [..] and customize it in a fairly easy manner." One of the analysts on the Dust 514 side said that for most of the people who they want to use the data "[w]e will need to deliver the results to people rather than expect them to go digging after them themselves, and at the same time find ways to encourage them to dig into the data themselves." An option would be to equip each team with an empowered individual whose responsibilities include data gathering and visualization. One of the users argued that this would make the feedback loop much shorter and enable them to get information very quickly, but the counterargument from another user was that this would indicate that the tools are simply too complicated to work with. The question is therefore if and how big data technology can help with data serving, accessibility, and ease of use.

Then there is the danger of relying too much on data. One from the users groups on the EVE Online side said his opinion was that "we should use the data to make sure we have all the information available when we make the decision", but when it comes to game development they would not be making "data driven features." Another one somewhat echoed this sentiment by saying his education and experience

[8]Targit is business intelligence & analytics software which can connect to OLAP cubes. Further information available at `http://www.targit.com/`.

Challenges of adopting big data in CCP
Improve analytical processes and formalise pipelines
Figure out how and what to report and how to turn information into actions
Define questions that require big data technology
Analytical resources currently insufficient
Proposed new projects represent massive undertakings and require new expertiese
Big data technology will becomes an independent resource consuming project
Get people excited about data and give them easy to use tools
Serve more data to people
Don't let data take over people's work
Balance data reliance and creative work
Make sure analysis is done well from the start in Dust 514

Table 7: Challenges of adopting big data in CCP

made him strongly favour being data driven and focusing heavily on using that as a guidance, but it would be dangerous to focus too much on data driven decision making everywhere because then you run the risk of becoming too introverted and not expanding the marketplace. In his opinion it is very important to have the creative people who can look beyond the data because he relies heavily on "the creative part of the company to open up new avenues that I can then take and optimize and make better." He went on to say that data could of course open up people's minds and help them figure out what they want to do, but leaning do much on data would leave out "the creative part of the company, which is one of the greatest assets CCP has", and they therefore would need to strike a good balance between the two.

A compilation of the challenges CCP faces with adopting big data is presented in Table 7.

5.2.6 Who in CCP Should Use Big Data Analysis?

Most of the interviewees came to the conclusion that as many as possible in the company should be users of big data in the sense that they should either be using big data technology to perform analyses or be the recipients of the information that comes out of such analyses. One influencer said this would eventually be "so integrated into our systems and our processes that everyone that deals with our projects would be a user of it." The groups most often mentioned were Customer Support, Community Management, Research & Statistics, Game Designers, Producers, and Marketing, but the finance teams, upper management and the corporate level were also mentioned as eventual users of the information coming out of this. The above

influencer added that "in the end everyone that deals directly with decision making within the company at all levels" should be a user. Many additionally identified the players themselves as probably the key users of the information coming out of big data analytics.

The Research & Statistics (R&S) department handles a big chunk of the data analysis and serving within CCP. They have specialized analysts and were mentioned as the department that would "help people through this." R&S were described as the people who would do actual data mining and look for explanations in the data. People would come to them with needs and requirements that would then be turned into reports or tools that allow people to interpret, understand, and drill further. They would be the ones doing the "heavy lifting" and also be responsible for providing the teams or managers with recommendations based on their analyses. But R&S could also provide some agreed upon basic metrics that describe particular situations and then provide people with more detailed data if they want to dig further and have the knowledge to do so.

One of the managers said that in a perfect world he would like CCP to have a nice interface that any developer or employee could use to put forward a hypothesis, query the data and create a report to test their hypothesis. If the hypothesis or idea is too big then they would have to turn to R&S, but his dream scenario is one where anyone can query the data sets. Then people would come around when they see the available information, use it a bit and think to themselves "it would be good if I have X" and that way get involved. The challenge there is to make sure people both fully understand the data they are looking at and that they aren't spending all their time just looking at data instead of doing their jobs. They may in some cases need to develop a certain skillset before they can perform these investigations, but the managers said they "see zero harm in equipping as many people as possible with basic skills to do some digging" to make them "self-sufficient" and thus ease the load on R&S and shorten the feedback loop. One of the influencers said that if CCP cannot get to that position soon "then at least every SCRUM team should be equipped to ask these questions, and every department should have the ability to look into this data for any kind of purpose that they are interested in." Another influencer added that if people are doing something wrong in their analysis, or abusing the data, then it would be their manager's responsibility to identify that, for example by comparing to the baseline metrics provided by R&S or by asking someone else to look at the analysis. He said he would prefer "an open system that has a baseline or benchmark and a peer-review process." Having this as open as possible would enable others to go in and fact check as well as facilitating healthy discussions.

The players themselves were regarded as perhaps the key users of the information coming out of big data analysis. The killmails[9] were mentioned as a good example of unstructured data that essentially became a product. The killmails were picked up by the players and harvested by corporations and alliances into killboards where they can track their kills and losses. A manager said that when CCP has more access

[9]When a player's ship or a structure belonging to players is destroyed a killmail is sent to the victim as well as the individual who delivered the final blow. It includes details on the victim's ship, its fittings, cargo contents, and what items were destroyed and what dropped in space, who was involved and how much damage they inflicted.

to data through the big data project it could start "developing products around that data" and present them to anyone who either plays the game or is interested in it. The transactions, kills, political drama, territorial wars, and everything that goes on in EVE Online could thus become content for both the people generating it as well as anyone else. Within the right setting and framework this information could be so valuable and interesting to players that it could be sold to them, thus making them customers of CCP's big data efforts.

Game Designers and developers were mentioned as one of the main users of big data analysis in that it could help them find the thing that "really makes EVE Online tick, because it is a very complicated thing to put your finger on." There are not a lot of people who have a clear understanding of this "thing", but it would be immensely helpful if they could query the blob of data to help them figure it out. They could also use it to answer all sorts of ideas they might get and measure the features and changes they put out. In short, the interviewees said it would help developers better understand how people are behaving in the game, and the sooner CCP can get to that point the better. With EVE Metrics the game designer's view has already become that it is "unrealistic from this point on that we do features that don't come with a stat tracking set."

Marketing was mentioned as needing to be a big user of big data in order to be better able to understand purchase behaviour, all the way from VIP subscriptions to marketing campaign performance. A virtual sales team was also mentioned as a big user of this because they would be looking in real time at what is being purchased and why, and they need to understand it so they can react to it directly.

Customer Support and Community management would also be important users. Customer Support personnel need access to aspects of big data to make their jobs easier, but they would be perfectly capable of operating with a fairly limited set of the data. What is important for them is to get a clear view of as much as possible of player actions through detailed logs, and the ability to pivot the data and slice it in their investigations. Community Management was mentioned as requiring big data analysis in a very large way because they are "part of the customer lifecycle management and how we can optimise sales to those users." The higher ups in the company, the managers, producers, and the people who run the business were also mentioned as needing good information to do that. People should not be spending all their time looking for and trying to understand data because it needs to be "as easily digestible and easy to get as possible so they can make good decisions from it."

In short the view is that as many as possible should have easy access to as much data as they need to make informed decisions in the time it matters and to help them understand the various aspects of the games.

6 Discussion

CCP benefits from many of its staff on all levels having a strong technological background, and the understanding of what big data offers is certainly there. Several current challenges were presented that could be overcome with the help of big data technology, and the new opportunities discussed were very ambitious. CCP is not a company that has lacked ambition or eagerness in the past and has always strived for excellence, but in the circle of build-measure-learn the latter two have been suffering while the building has progressed at full speed. CCP suffered a significant business setback in late 2011 and as a result the company went through extensive restructuring where it reformed around its core business and set out to be more data driven in the way it operates. Even before this shift there were people within CCP who felt for a long time that the company was moving towards having to resort to new solutions to solve its looming data challenges with both EVE Online and Dust 514. Of the people interviewed, nearly all were convinced that adopting big data technology was the best step forward to tackle these challenges. In terms of need it was clear to most that Dust 514 and its free-to-play (F2P) business model requires this and that EVE Online, while not requiring it at the moment, could benefit greatly from having this analytical capability down the road. The competing MMO developers and publishers are largely moving their subscription based games over to the F2P model and many of the new and upcoming MMO games are built from the ground up with F2P at their core. It is a tough market and the next years are likely to see hard competition for each subscriber and each dollar spent on games like this. Players have the ability to very easily switch between F2P games if they don't like something about the game, and with so many offerings they are likely to want to spend their money very carefully. With a subscription game such as EVE Online you have some lead time to react to a situation. Once people have committed to a long-term subscription it takes months for them to leave it so CCP has much more leeway there. In a F2P game you have a much shorter lead time because if your items are not selling then you are not making any money, and if people don't understand or enjoy the game from the beginning then they are unlikely to return. It is the opinion of this author that, even though the need does not seem so critical at the moment, CCP should equip itself with big data technology and the associated analytical expertise as soon as possible, not just because of Dust 514 but to ensure they maintain their competitiveness and are not left behind in the MMO market.

During the interviews the technology and the opportunities it presents for CCP were spoken of in a somewhat gung-ho manner by some. This is perhaps nothing to be alarmed about given the casual interview setting, but CCP must investigate whether it truly has the capacity for this type of work. Those already working on the analytical side had a more serious outlook because they know first-hand that CCP still lacks the resources, tools, and processes for fully utilizing the already available data and implementing the available projects. Getting big data technology on the table will not solve the resource problem and is rather likely to increase the number of requested projects. Just figuring out what kind of resources CCP needs in order to incorporate big data analysis into their system is a whole project by itself. The more ambitious ideas, such as recommendation engines, will further require expertise that CCP currently does not have, and the requirements put forward for Research

& Statistics to fulfil will require that department to grow by several data specialists, including experienced big data developers.

Specific cases can be solved using current technology to satisfy logging and analytical needs, but CCP needs a setup where this is so cheap in terms of logging, computations, and storage that they can start doing it for as much of the game as possible. It is important for CCP to gain an understanding of what sort of problems the currently available technology (SQL Server, Excel, PowerPivot, Targit, etc.) is best suited for solving versus where big data technology is necessary. People are used to the current technology, it represents their comfort zone, and they are likely to want to use that technology so solve even those problems that cross over into big data territory. That would be unacceptable as the solution might end up being too big and cumbersome to use and certainly would not scale well. In the same way a problem solved using big data technology when the current technology would have been much better suited would end in an over-engineered solution where they kill a fly with a cannon. As the experience of using big data technology grows it will become easier to sort problems depending on the problem domain and therefore it is important to get started with assigning people to icebreaker projects to get familiar with the technology and processes involved. These could be projects where they can get some early wins such as with Customer Support or feeding information back to the player base, or they could be entirely experimental projects with no benefit other than looking good and helping the developers learn, but they would preferably include someone experienced with the technology, in this case Hadoop.

CCP is challenged with internal usage of data for feature development, customer service, business intelligence and many other purposes. These challenges sound like nothing out of the ordinary for a regular business, but when you have an enormous and complicated system such as EVE Online with nearly 400 thousand active players then these needs quickly go beyond being ordinary. The managers believe that in order for big data solutions to become successful they need to become parts of the development service platform and must enable the same level of ease of use and access for the teams as EVE Metrics. This needs to happen not just on the big data level but also with the analytical solutions built using the current technology. This is very important if the intent is to strengthen the build-measure-learn process, but it also means that many developers' attitudes towards data utilization must improve. EVE Metrics has done wonders for some people, but data usage needs to be incorporated on as many levels and in as many teams as possible for this to be successful. If some developers still resist using EVE Metrics or other currently available data solutions to help their work then it is doubtful they will be interested in using big data solutions. Getting up to speed with big data is complicated and resource intensive and will be useless if people are not interested in using the information coming out of it. This points to the clear need for good and intuitive tools for those who want to dig deep in the data as well as those who simply want to take closer look, and the importance of reports, dashboards, and other forms of delivering information to people that will pique their interest. Research has shown that the education system has not prepared people well for analysing data, even when it comes to something as simple sounding as explaining graphs[14]. The same research showed that it is ineffective to train people to focus on the analytical tools and instead they should be "taught how to interact with the data" and think

critically [14]. CCP needs to educate its people in data usage, starting as soon as possible.

Continuing on this tangent, the problem has never been that CCP had too little data but rather that CCP was not processing the data it had. Over the last 5-6 years they could have been doing much more with the data, such as letting game design be influenced to a much greater extent by the logged behaviour of players. CCP did not need big data or EVE Metrics or any new systems to be able to do that at the time because they already had millions of log events sitting unused on the server. A feature could be added to the game and a year later no one had a clear idea of how it was performing. All the data about it was generated on the server, there was simply no one asking questions about it. Developers cannot say that they did not know how a feature was performing for a year because they did not have Hadoop, they simply were not looking. The main thing now is this awakening to start using the data. CCP is not "pressing the big data button", they are adding new tools to their toolset in order to be better at analysing data and doing their jobs. The challenge is, and has always been, how CCP can use knowledge of what goes on in the game world to improve their games and their service. Having more logs and new technology does not change that, it empowers people by giving them more data, better tools, and new opportunities. The restructuring brought about this new emphasis on data and actionable metrics, and CCP has to ensure it gets serious on this issue because even though they start storing orders of magnitude more data nothing happens if no one is using it.

As much as this conclusion has emphasised the need for CCP to plan for and follow through with the challenges presented by both adopting big data technology and increasing the role of data in decision making across the company, it is important to always have data-sceptics on board who can push back against the analysts. The idea that complicated algorithms and number crunchers should be making all the decisions, solving all the problems, and leading the company, is a dangerous one. It is important to separate the analysts and those who have the experience and talent to lead the company. This is simply a precaution because the decision making should be aided by data wherever possible, especially in very competitive markets, but the leaders and others in the CCP need to operate as 'informed sceptics' "who rely on data but not so much that they are afraid to question the results and solicit feedback from others."[14] This has a few consequences:

- The analysts within CCP must be willing to and capable of explaining their methods and results.
- Starting from the top, the managers must improve their abilities to understand and evaluate information. This has the potential to influence others and carry down the chain.
- Employees need to be empowered to ask questions and seek answers themselves.
- Data efforts must be focused on obtaining actionable information.
- As much as possible of the features already in the games, and certainly all those added, must have clear metrics associated with them.
- Data needs to be taken seriously but also balanced with people's intellect and insights.

CCP has a company culture of openness and transparency. It has shown its ability to adapt and change the way it operates. Everyone in the company is enthusiastic about what they are creating and want as many of the players as possible to enjoy it. They just need to be more active when it comes to measuring and learning from what they have built.

6.1 Recommendations For Dust 514

With the sheer amount of users predicted to participate in Dust 514, the different sever structure, and the free-to-play revenue model it is clear that the technological solutions and processes used for data extraction and analysis in EVE Online will not be sufficient for Dust 514. While Dust would be able to survive without big data technology for a short time after launch, the pressing need to fine tune the game experience and revenue model will require big data technology as soon as possible.

The Dust development team has compiled a large list of information and metrics they must have, defined the type of data mining that needs to be done, and created numerous mock-ups of dashboards and reports with that information. As explained, it is very important for them to have as much of this as possible when the game launches to support the business model, and that hinges on having the right frameworks and pipelines in place. Work needs to start as soon as possible on defining and building those frameworks and putting the required technological pipelines in place as well as the professionals needed to fit all the pieces of the puzzle. The high-level understanding and urgency is all there, they just need to get boots on the ground.

6.2 Recommendations For EVE Online

EVE Online has employed a successful formula and continually expanded its player base for nine years without using big data technology. There is no perceived pressing need at the moment for the EVE Online developers to start using big data technology to solve their problems, but the EVE Metrics project has gotten the ball rolling and managed to make significant advances in getting the developers more data hungry. It is critical to keep this momentum going and use the opportunity to start training people in data usage in line with the company's new emphasis on being data driven. It is already known that in 12 to 18 months they will absolutely need this from their people so preparations should start immediately.

7 Conclusions

In today's highly competitive business of MMO games, companies will stand or fall with how they leverage the data they collect. CCP hf is presented with many opportunities and hard challenges in its planned adoption of big data analysis. The expectations are already set very high, but to achieve them CCP needs to change the way it operates. The way CCP thinks about and approaches data must evolve and mature over the next few years and become fully integrated into how they operate both EVE Online and Dust 514. In the short term they need to start equipping themselves for this as soon as possible by changing the way their staff approaches, thinks about, and is served data. Data training of managerial staff and other decision makers is necessary, as well as hiring more analysts and data scientists. The successful adoption of big data and data science at companies has transformed the way they operate in a way that puts these processes at the core. It is likely that this will happen with CCP as well, but they need to ensure they end up in a place where they gain the best of both worlds with their creative element and data analysis.

A Interviewee Details

A.1 Management

Hilmar V. Pétursson - CEO
Hilmar is the Chief Operating Officer of CCP. He sets the top level business strategy for CCP.

Jón H. Jónasson - COO and CCP Asia Manager
As Chief Operating Officer, Jónasson's responsibilities cover teams such as Web Devlopment, Office IT, Customer Support, Community Management, and Research & Statistics. As Manager of CCP Asia he is in charge of the Shanghai office's overall operations which is responsible for a bulk of Dust 514 development as well as Customer Support for EVE Online in China.

Halldór Fannar - CTO and PO
As Chief Technology Officer, Fannar's primary focus is on the technological challenges facing the company and the long term-term issues. Fannar is also a Product Owner[10] within the EVE project, putting some of the product development parts of EVE Online under his direction. Fannar also has some people management responsibilities.

Jon Lander - Senior Producer of EVE Online
As Senior Producer, Lander is responsible for everything relating to EVE Online.

Brandon Laurino - Executive Producer of Dust 514
As Executive Producer Laurino is responsible for everything relating to Dust 514.

A.2 Influencers

Andie Nordgren - Technical Producer for EVE Online
Nordgren is a Technical Producer for EVE Online in the accessibilty segment where she works with the EVE cluster team and the shared services team. She is also an influencer when it comes to how CCP can improve its strategic planning.

Eyjólfur Guðmundsson - Lead Economist and Director of R&S
Guðmundsson is the director of Research & Statistics, a department whose responsibility is to serve CCP with accurate and actionable information on player behavior in the virtual societies operated by CCP. He is also the Lead Economist for CCP responsible for providing information on the real and in-game economies of CCP Games, handles academic relations, and is involved with various initiatives inside the company.

Jón Bjarnason - Programmer
Bjarnason is building the eventlog system and the EVE Metrics initiative, and will take partial responsibility for integrating Hadoop into CCP's processes.

[10] SCRUM terminology. The PO manages a product backlog and prioritization of the items within.

A.3 Users

Ingólfur V. Ævarsson - Director, Customer Lifecycle Management
As the director of Customer Lifecycle Management, Ævarsson is responsible for numerous analysis in user behavior, both customer and non-customer. He then works with various teams to apply results of the analysis to optimize processes and features to increase customer acquisition, retention and re-acquisition.

Kristoffer Touborg - Lead Game Designer for EVE Online
As Lead Game Designer Touborg is responsible for the health of the EVE Online gameplay, developing new features for expansions, and running the development teams.

Eino Joas- Game Designer for Dust 514 and Dust 514 economist
Joas is responsible for spearheading the Dust 514 analytical processes, developing the game's virtual goods ecosystem and monetization strategy, as well as designing its virtual economy and how it links to EVE Online.

Ben Cockerill - Producer on Dust 514
Cockerill is a producer for the Virtual Goods strategy for Dust 514, sharing that responsibility with Joas and others.

References

[1] *A different game. Information is transforming traditional businesses.* Feb. 25, 2010. URL: http://www.economist.com/node/15557465 (visited on 02/21/2012).

[2] Chris Anderson. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.* Wired. June 23, 2008. URL: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (visited on 02/20/2012).

[3] Dhruba Borthakur. *Realtime Hadoop usage at Facebook. Part 2 - Workload Types.* May 28, 2011. URL: http://hadoopblog.blogspot.com/2011/05/realtime-hadoop-usage-at-facebook-part_28.html (visited on 03/05/2012).

[4] *Clicking for gold. How internet companies profit from data on the web.* Feb. 25, 2010. URL: http://www.economist.com/node/15557431 (visited on 02/21/2012).

[5] *Data, data everywhere.* Feb. 25, 2010. URL: http://www.economist.com/node/15557443 (visited on 02/25/2012).

[6] *HDFS Architecture Guide.* Apache Software Foundation. 2012. URL: http://hadoop.apache.org/common/docs/current/hdfs_design.html (visited on 04/22/2012).

[7] Steve Lohr. *For Today's Graduate, Just One Word: Statistics.* Aug. 5, 2009. URL: http://www.nytimes.com/2009/08/06/technology/06stats.html (visited on 02/25/2012).

[8] Steve Lohr. *The Age of Big Data.* Feb. 11, 2012. URL: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html (visited on 02/25/2012).

[9] Mike Loukides. *What is data science?* O'Reilly. June 2, 2010. URL: http://radar.oreilly.com/2010/06/what-is-data-science.html (visited on 02/20/2012).

[10] James Manyika. *Big data: The next frontier for innovation, competition, and productivity.* Research Report. McKinsey & Company, May 2011. URL: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation (visited on 02/25/2012).

[11] *MapReduce: Simplified Data Processing on Large Clusters.* HTML Slides. Google. 2004. URL: http://research.google.com/archive/mapreduce.html (visited on 17/02/2012).

[12] Cade Metz. *How Yahoo Spawned Hadoop, the Future of Big Data.* Wired. Oct. 18, 2011. URL: http://www.wired.com/wiredenterprise/2011/10/how-yahoo-spawned-hadoop/all/1 (visited on 02/20/2012).

[13] Kamesh Pemmaraju. *Drowning in "Big Data" Noise: Where's the Real Signal?* Sand Hill Group. Oct. 3, 2011. URL: http://sandhill.com/article/drowning-in-%E2%80%9Cbig-data%E2%80%9D-noise-where%E2%80%99s-the-real-signal/ (visited on 02/17/2012).

[14] Ethan Rouen. *Big Data won't solve your company's problems.* CNN. Mar. 19, 2012. URL: http://management.fortune.cnn.com/2012/03/19/big-data-wont-solve-your-companys-problems/ (visited on 04/28/2012).

[15] Ashish Thusoo. *Data Warehousing and Analytics Infrastructure at Facebook*. Tech. rep. Facebook, June 6, 2010. URL: http://borthakur.com/ftp/sigmodwarehouse2010.pdf (visited on 03/05/2012).

[16] Ashish Thusoo. *Rethinking the Data Warehouse with Hadoop and Hive*. Cloudera. Oct. 2, 2009. URL: http://www.cloudera.com/resource/hw09_rethinking_the_data_warehouse_with_hadoop_and_hive/ (visited on 02/27/2012).

[17] *Unit sales of hardware (since April 2006)*. Sony Computer Entertainment Inc. 2012. URL: http://www.scei.co.jp/corporate/data/bizdataps3_sale_e.html (visited on 04/23/2012).

[18] *Welcome to Apache™Hadoop™!* Apache Software Foundation. 2012. URL: http://hadoop.apache.org/ (visited on 04/22/2012).

[19] Tom White. *Hadoop: The Definitive Guide*. Ed. by Mike Loukides. Second. O'Reilly Media, 2011.