DECISION SCIENCES INSTITUTE
A Proposal to Improve Patent Search with Big Data Analytics

Steven Manns
Citibank
Email: stevewmanns@gmail.com

Richard J. Goeke
Widener University
Email: rgoeke@widener.edu

ABSTRACT

Firms increasingly recognize that an IP portfolio represents an important competitive advantage, so patents have shifted from a back-office function into an important and strategic component of business strategy. Obtaining a patent, however, is still a difficult and time-consuming process, because an innovation must prove sufficient 'novelty'. This is made all the more challenging today, because patents are being filed, sold, and traded in record numbers. We contend that IP and patent search problems can be addressed by readily available big data and analytics technologies, and offer a workable plan to move forward.

KEYWORDS: Patent Search, Intellectual Property, Big Data, Analytics

INTRODUCTION

Intellectual property (IP) is an asset that has attributes similar to tangible assets, in that both can be bought, licensed, or sold. In order to protect the rights to an IP asset, a company will typically file a patent, which must prove that the invention be novel, unique, useful, and nonobvious, while also being capable of some practical application. In return, the inventor is given an exclusive right that prevents others from using and profiting from the invention. These protections, and the profits that often come with them, have made the IP and patent marketspace a big business. An entire new economy has emerged, in which patents are licensed, traded, auctioned, donated, sold on exchanges, and even strategically held to prevent competition from entering marketspaces. There is a lot at stake for many companies, both large and small. This modern-day gold rush has companies clamoring to find new ways to gain marketspace dominance.

A wealth of patent information exists in public search domains, such as the U.S. Patent and Trademark Office (USPTO) database and the PACER system, which houses all federal court records. Internationally, public patent information from Europe and Asia-Pacific countries are aggregated and made available via web search. Other databases exist, such as the NIH PubMed that houses over 26 million searchable research publications on biotechnology. Tracking new patents and leveraging these databases via effective search techniques can be a source of competitive advantage (Marydee, 1993). But the problem is that traditional methods to search these databases are antiquated, because they rely on keyword searches using Boolean search technology (e.g. AND, OR, NOT). Although this type of search methodology does work, tests have demonstrated that only 22 to 57 percent of relevant documents are located when using Boolean search (Waxer 2010). Often the process is very time-consuming, because many iterations of searches are required in order to hone in on the subject matter and

eliminate false positives. Because patent searches are so time-consuming, companies often employ trained experts, who can locate relevant documents more quickly, but this also makes the search process that much more expensive. In the end, there is always a long list to sort through with very little way to differentiate which items are more relevant than others (Sabroski 2012).

We believe that the present IP and patent search market can be revolutionized with big data and analytics. Through advances in hardware, storage, and software, big data and analytics has already pushed the traditional analytical boundaries by offering new insights into what was once deemed as not possible. For example, Williams-Sonoma used big data and analytics to improve the effectiveness of its merchant teams, by developing algorithms that efficiently and accurately match customers with products and cross-selling opportunities (Alber 2014). In the patent litigation area, a solution has been developed that help companies and lawyers predict the likely outcomes of patent litigation cases. If a company is facing an infringement lawsuit, Lex Machina can be used to determine the likely outcome of the case and help determine if the case should be settled outside of court or not (Ard 2016). We argue that big data and analytics can be applied to publicly available patent data, to improve patent searches, obtain relevant results more quickly, and obtain a single holistic picture.

LITERATURE REVIEW

In today's economy, a firm's IP and patent portfolio has become a tremendous source of firm value, but these intangible assets are under near continuous threat (Sabroski 2012). To get a sense of how important IP and patents are to the U.S. economy, it has been estimated that of the combined $17.9 trillion market capitalization for the S&P 500, more than $8 trillion has been listed on their balance sheets as intangible assets (Monga 2016). And, the U.S. is not alone in the IP and patent marketspace. China went from 40,000 patent applications in 2003 to 600,000 patent applications in 2013 (Reuters 2014). Companies today are in a high stakes competition to quickly innovate while trying to also anticipate their competitor's next move. The stakes are much greater for smaller companies, because IP assets make up a much larger portion of the small firm's value, which is generally used to attract private funding (Steeves 2015).

As if the stakes were not already high for companies to innovate, a new breed of innovation companies has cropped up - 'Patent Trolls' have been spawned as shell companies that acquire undervalued patent rights. The patent troll companies "can exploit their assets by bundling patents together and may ultimately act as entities that exist solely to pursue infringement cases" (Abril and Plant 2007). In 2015, venture capitalists invested $18 billion in patent shell companies, and another $7 billion in the first quarter of 2016. Patent trolls have had a substantial (and often negative) impact on innovators and startup companies – in 2015, a record 7,500 patent disputes were filed (Sole 2016). This patent war has taken a toll on may industries, not just in the technology sector. As a defense, companies have banded together, by licensing and sharing patents within a private market or network. These networks (referred to as 'Patent Thickets'), can help companies reduce their overall risks (Sole 2016). But the value of and necessity to properly manage the firm's patent and IP portfolio has never been greater. If the firm's IP and patent portfolio can be properly managed, then substantial profits and other advantages can be gained. A recent study concluded that "any company that owns more than 450 patents and spends in excess of $50 million on research and development should be able to generate between 5%-10% of its operating income from its IP assets" (Abril and Plant 2007). In short, patents are the lifeblood of almost any company today (Sabroski 2012), which demonstrates that companies need to navigate the complex web of patent information to remain competitive.

The tremendous volume and complexity of patent information available makes obtaining a patent all the more difficult, and this has tremendous implications for the patent application process. Not only is there a record number of patents filed each year, but there is large amount of knowledge and information in and around patents. This presents a significant problem when trying to file for a patent, because of the novelty requirement. In other words, if there is prior-art that describes the invention in the form of a publication, lecture, book, or other content, then the patent application can be rejected. Even down the road, if prior-art is not thoroughly researched, an opposing party can challenge the patent and have it invalidated (USPTO website, 2016). Therefore, it is critical to perform a through patent search, but performing an exhaustive search on a patent is an iterative, time-consuming, and expensive process. A search must be repeated many times across many information sources, and is like trying to find the proverbial needle in the haystack, but doing so on a routine basis (Sabroski 2012).

To conduct a patent search, most patent sources (e.g., USPTO in the U.S., WIPO in Geneva, Switzerland) provide basic tools that use Boolean methods (e.g. AND, OR, NOT) via a web interface. However, there is a strong consensus that keyword Boolean search is a failure and can lead to mispresented results (Sabroski 2012). Sabroski (2012) notes that a semantic search, which converts search terms into contextual meaning, can improve search results, and firms such as LexisNexis and Thomson Innovation have developed tools that make use of higher level semantic search methodologies. A different approach was suggested by Koch and Bosch (2011), who proposed a system that could seamlessly perform a consecutive and iterative query refinement process on the data. The refinement process is managed by the user through visual exploration techniques. However, the authors noted that there are scalability limitations with their proposal. Unfortunately, neither Boolean search methods nor those offered by private firms are well suited for today's complex patent data, which we suggest is a classic big data scenario.

We argue that patent data is a classic big data problem, defined in terms of the 3 V's - volume, variety, and velocity. Patents are very dense with a lot of information and many attributes. Patent density is also attributed to forward and backward references to other patents, citations to external research, and references to legal events - all referred to as patent linking. Linking is similar to a hypertext link within a HTML page. A patent can contain several links to other patents (forward citations) and other patents can reference the patent (reverse citations) (Wang 2015). Patent linking across 50 million patents that date back to as early as 1617 is a big data volume problem. The World Intellectual Property Organization (WIPO) is an international organization that covers the Patent Cooperation Treaty and patents from many counties, including the USPTO, in a database called PATENTSCOPE. As of November 2016, the WIPO PATENTSCOPE database reported 50 million searchable patent documents. The second big data V, variety, fits the patent analysis problem well. Patents extend across many industries and intellectual property spaces. There is a patent in almost every subject, from very small molecules found in nanotechnology to complex systems installed in large manufacturing facilities. Variety also refers to the diverse amount of information found within a patent. Patents can include drawings, images, diagrams, flowcharts, and all of the above. Finally, the third big data V refers to the velocity that the data is generated. The patent system is based on the concept of novelty. A patent will not be granted to an applicant if there is prior knowledge, also known as prior art. Prior art not only refers to prior patents, but also any information that exists in the public domain. Therefore, all public information in print form is considered prior art. De Wachter (2013) argues that big data is prior art, and therefore, anything that involves patent analysis is involves big data. The velocity of big data, in addition to the tremendous growth in

patents filed, demonstrates the dire need for a robust and scalable patent analysis methodology (De Wachter 2013).

There is a significant gap between current patent search capabilities and what businesses need in order to be effective. Companies need a holistic patent analysis approach in order to achieve competitive intelligence and IP portfolio management. The focus of this research is based on the methodologies and algorithms that are being used in other areas and that can easily translate to patent analysis. The end solutions being developed in the public and private sector are not in the scope of this paper. However, patent data is public data that is freely accessible in more than 100 patent offices worldwide. And, the very premise of a patent is protection for the patent holder, along with the dissemination of information and knowledge, which can be used to either improve upon an invention or lead to an entirely new invention. Therefore, the focus of this paper is to identify a big data methodology and system that can be used to advance patent analysis in the open source arena. In this regard, not only can companies with deep wallets benefit, but also entrepreneurs with more limited budgets.

PROPOSED SOLUTION

Given that patents and IP seem to meet the definition of big data, then a big data analytics solution seems appropriate. In fact, the Open Source Patent Analytics Project has already started down this path (Oldham 2016). The project thus far has provided an overview for converting patent data into a big data analytics project, but little progress has been made since April 2015. By comparison, the goal of this paper is to provide more detail regarding the process of transforming patent data into a workable big data and analytics solution that can be used by large companies and small business entrepreneurs.

Before any big data analytics methods can be applied, the first step is to extract the data from public patent sources and load it into a scalable petabyte storage system. An extraction process for patent data could easily be modelled after other similar extraction processes. For example, the National Center for Biotechnology Innovation (NCBI) publishes large genomics databases and research related information on PUBMED. The well-established field of bioinformatics has developed open processes and tools to extract and load large genomics datasets from public sources. Once the data has been extracted, informatics tools are used to parallelize the computational work across large compute clusters. Often the public bioinformatics data is merged with private genomic sequences. This same model can be adapted to extract patent information from USPTO, WIPO, and other public data sources. Recently, the USPTO published a series of application program interfaces (APIs) that now offer seamless integration (Spence & Bijman, 2016). Specifically, one can set up batch processes to download new patent data via the PAIR Bulk Data API. Further details can be found at developer.uspto.gov/api-catalog. WIPO has similar API capabilities for international patents.

Once the data has been extracted, it can be merged with internal IP information, then loaded into a highly scalable and fault tolerant Hadoop Distributed File System (HDFS) or another objected-oriented based filesystem hosted in Amazon Web Services (AWS). A lot of the bioinformatics work has shifted to HDFS, which is the underlying filesystem for Hadoop. Also, AWS is a cost-effective means to quickly provision a 50 node Hadoop cluster in less than an hour. The HDFS filesystem uses a uniform namespace that is visible from any node within the Hadoop cluster. Files are stored in HDFS and other object based systems in a manner that is similar to a key value pair. Every patent is uniquely identified by a publication number. Therefore, each patent will be stored in the Hadoop filesystem with the publication number as the key. The key will be preceded by the name of the data source, such as WIPO/PatentABC or

USPTO/PatentXYZ. For example, a patent number of 1234 filed with the USPTO will be stored as USPTO-1234. Patents are automatically linked to various data sources because there are attribute fields in the patent that reference other sources. For example, a USPTO patent may have an attribute field that references a WIPO publication number. Therefore, everything is cross-linked between data sources. For smaller data sets, one could spawn an AWS data lake to store the patent data in a similar manner. There are many storage options available in AWS. Appendix 1 shows the steps to spin a Hadoop cluster in AWS with Hive or HBase. After the cluster has been fully provisioned, Python scripts can scrape the USPTO and WIPO sites for changes. The scripts can pull the new data down from the remote sources and merge it into the Hadoop cluster on a desired interval using a job scheduler.

Once the cluster has been created in AWS, the next part to the big data analytics solution is to process the data. Data can be processed many ways, and the tools we're proposing come from the open-source world. This is not to say that other tools (e.g. Microsoft, SAS, Tableau) cannot be used, but these tools carry a licensing fee. Rather, we are proposing the use of open-source tools, such as R, RStudio, Shiny Server, and Python. R is a programming language for statistical analysis and graphics., RStudio extends an IDE (integrated development environment) to assist in the development of R applications, and Shiny integrates with R as a web framework to graphically display the analysis. These three tools (R, RStudio, and Shiny) provide a powerful statistical analyses environment, including a user-friendly IDE on a fully managed Hadoop environment that starts up in minutes, and saves time and money for data-driven analyses (Schmidberger 2014). Whether the task is a simple text analysis across patent databases or a large pipeline analysis using natural language processing and machine learning capabilities, deploying a Hadoop cluster with R on AWS is the foundation to achieving big data analysis at an economical expense. As the needs grow or shrink, the cluster can easily be resized to meet the analytics needs.

The last part of our proposed solution is the visualization of the data. Data visualization is very important to help aid in the interpretation of result sets. There are open-source and licensed options that can be directly linked to the back-end AWS storage. The Shiny framework (open-source) can be deployed on the head node of the Hadoop cluster and a web service can serve up the data dashboard to the end-user. Appendix 2 shows an image of a Shiny dashboard that is streaming live data from a site's download logs to depict the popularity of each application type in real-time. Another option in AWS is Tableau. In the bottom frame of Appendix 2, a treemap groups patents published in each technology category. Tableau can easily be added to an AWS instance and linked to the S3 data store very little cost, and can be shut down when not in use to prevent additional billing.

DISCUSSION AND CONCLUSIONS

Deploying an IP and competitive intelligence strategy does not need to be a complex or expensive undertaking in the new world of big data and analytics. Methodologies from fields such as bioinformatics and computer science can be adapted and utilized in the patent analysis space. Other technologies, such as machine learning and natural language processing, can easily be plugged into the analysis pipeline as well. It is easy for a general technologist to deploy a Hadoop instance in AWS and integrate analysis tools with visualization front-end dashboards. Scalability is no longer a factor. The Hadoop ecosystem, either deployed in AWS or in-house, can easily grow on cheap commodity hardware as the analytics needs expand. It is fairly easy to load petabyte datasets into Hadoop and run a distributed analysis using R or SAS in the Amazon cloud. Technology and money is no longer a limiting factor in achieving big data analytics in the patent world. The main limitation lies in expertise and the know-how to integrate

a full end-to-end big data analytics platform for patent analysis. Notwithstanding some of the high-profile successes that reported by some firms (e.g. Alber 2014), expertise with big data and analytics technologies is still emerging in the business world. Therefore, the chief obstacle to implementing the steps proposed here concerns the expertise of the people involved.

In conclusion, the solution presented here provides a workable plan for moving patent data into the big data analytics space. By using readily available technology, budget and scalability are no longer a limitation. And, by borrowing processes and frameworks from the biotechnology and bioengineering spaces, we have greater confidence that the methods proposed here will actually work. The patent analysis field has been slow in moving from legacy methods to big data methods. Big data can deliver on the IP business needs and close the technical gaps that currently exist. Open collaboration between the stakeholders in the patent space will help drive further advances in patent analysis. The challenge is to develop the expertise needed to implement this big data and analytics solution, and to move patent analysis to be more collaborative as in the research sciences. When these limitations are overcome, we believe that the path presented here will propel patent analysis into big data analytics.

REFERENCES

Alber, L. (2014) "How I did it: The CEO of Williams-Sonoma on blending instinct with analysis", *Harvard Business Review,* 92(9), 41–44.

Abril, P. & Plant, R. (2007) "The Patent Holder's Dilemma: Buy, Sell, or Troll?", *ACM Journal,* 50(1), 37-44

Ard, C. (2016) "The Rowdy Crowd, Disruptive Technologies in the Legal Industry," *Online Searcher Journal*, 14-18.

Bell, P. (2013, May/June) "Creating Competitive Advantage Using Big Data", *Ivey Business Journal Online.*

Koch, S. & Bosch, H. (2011) "Iterative Integration of Visual Insights during Scalable Patent Search and Analysis", *IEEE Computer Society*, 17(5), 557-568.

Marydee, O. (1993) "The patent and trademark databases: What can they do for you?", *Today, Inc*., 10(5), 8-9.

Monga, V. (2016, March 21), "Accounting's 21st Century Challenge: How to Value Intangible Assets," *Wall Street Journal Online*.

Oldham, P. (2016, November) "Open Source Patent Analytics Project," Oldham's Blog. Retrieved from poldham.github.io/

Sabroski, S. (2012, "Searching, Waiting, and Hoping for Semantic Search," *Online*, 36(2), 21-24.

Schmidberger, M. (2014, October) "Statistical Analysis with Open-Source R and RStudio on Amazon EWR," AWS Big Data Blog. Retrieved from aws.amazon.com/blogs/big-data/statistical-analysis-with-open-source-r-and-rstudio-on-amazon-emr/

Sole, M. (2016) "Fintech Patent Trolls Can't Be Allowed to Win," *American Banker,* 181(F364), 1-1.

Spence, M., Bijman, M. (2016, June) "USPTO's New API Expected to Shake-Up the Patent Tools Market," Chipworks. Retrieved from www.chipworks.com/about-chipworks/overview/blog/uspto%E2%80%99s-new-api-expected-shake-the-patent-tools-market

Steeves, R. (2015) "Show me the money," *InsideCounsel*, 26(283), 16-17, retrieved from Wolfgram Memorial Library online.

Thomson Reuters whitepaper (2014, "China emerges as world patent leader," Thomson Reuters. Retrieved from thomsonreuters.com/en/articles/2014/china-emerges-as-world-patent-leader.html

USPTO Open Data Portal, USPTO API's, Retrieved November 2016, from developer.uspto.gov/api-catalog

USPTO Website. (2016, November), “About Patents”, United States Patent And Trademark Office. Retrieved from www.uspto.gov/patents-maintaining-patent/patent-litigation/about-patents

De Wachter, J. (2013, September), “Big Data and Intellectual Property”, Joren De Wachter Blog. Retrieved from jorendewachter.com/2013/09/big-data-intellectual-property-2/

Wang, Y. (2015) “Identifying competitive intelligence of collaborative intellectual property alliances: Analytic platform and case studies”, *Information Systems & e-Business Management*, 14(3), 491-505.

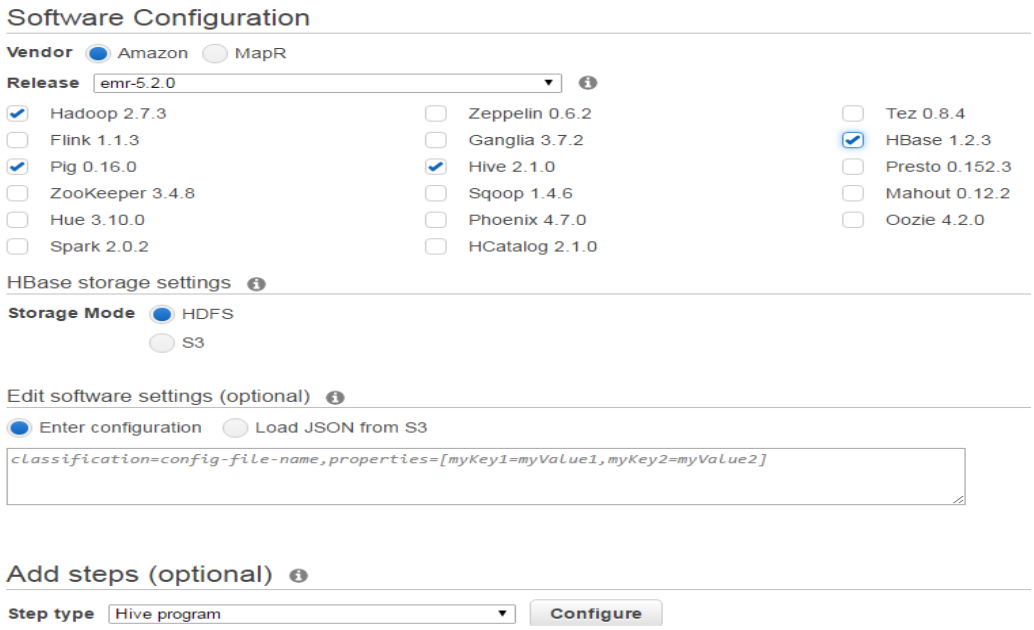
Waxer, C. (2010), “Keyword Searches Disappoint”, *Computerworld*, 44(23), 38-39.

WIPO PATENTSCOPE. (2016, November), Search International and National Patent Collections Database, WIPO. Retrieved from patentscope.wipo.int/search/en/search.jsf

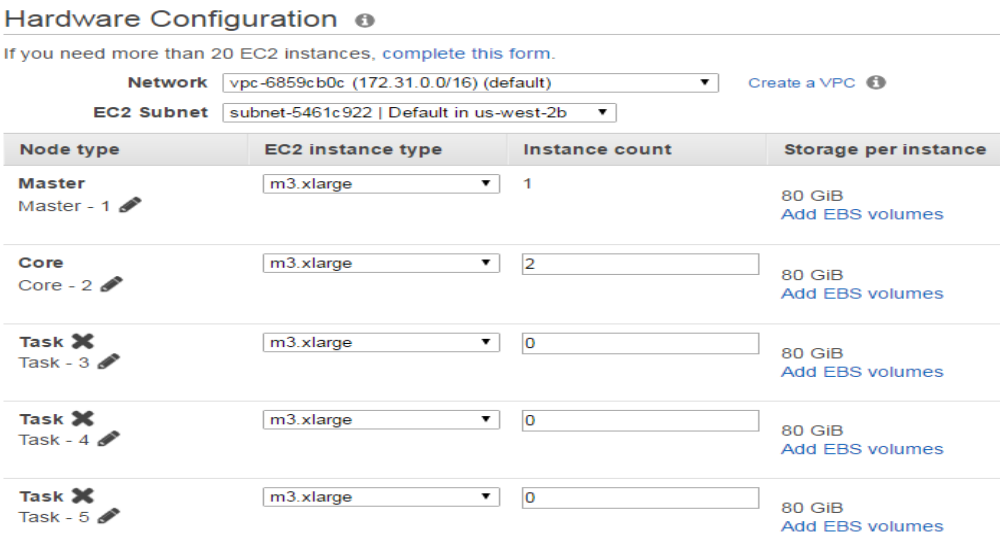
APPENDIX 1. Launching a Hadoop Cluster using Amazon Web Services

Step 1: Create the Hadoop Cluster
Go to http://aws.amazon.com

Step 2: Select the Software to Install with the Instance (Hive or HBase)
Hive sits on top of Hadoop and provides a SQL interface to insert and retrieve data

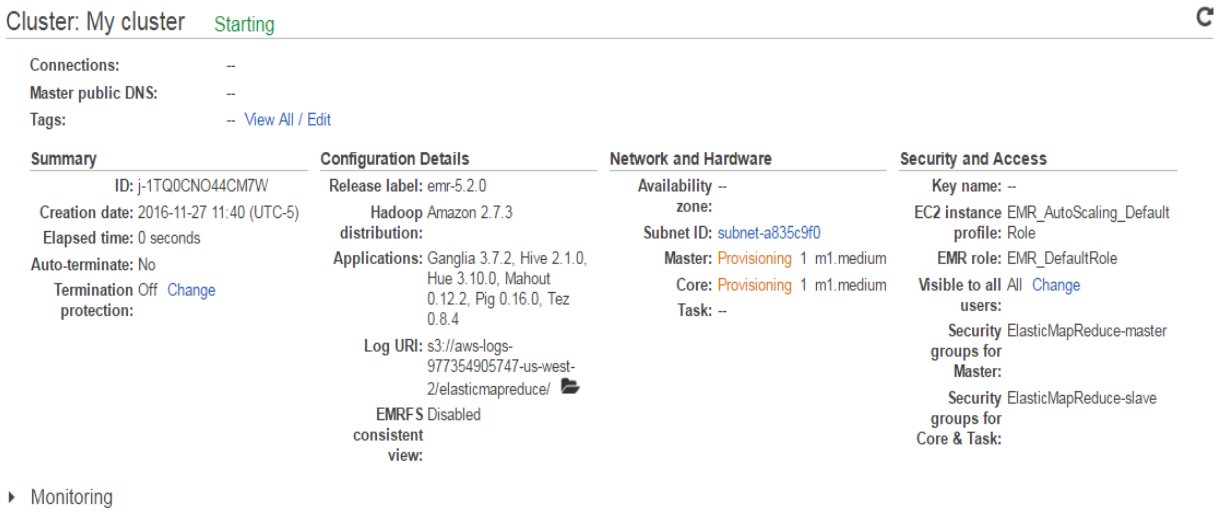


Step 3: Create the Nodes, then Select the Appropriate Node Configuration
Select the appropriate configuration for the cluster



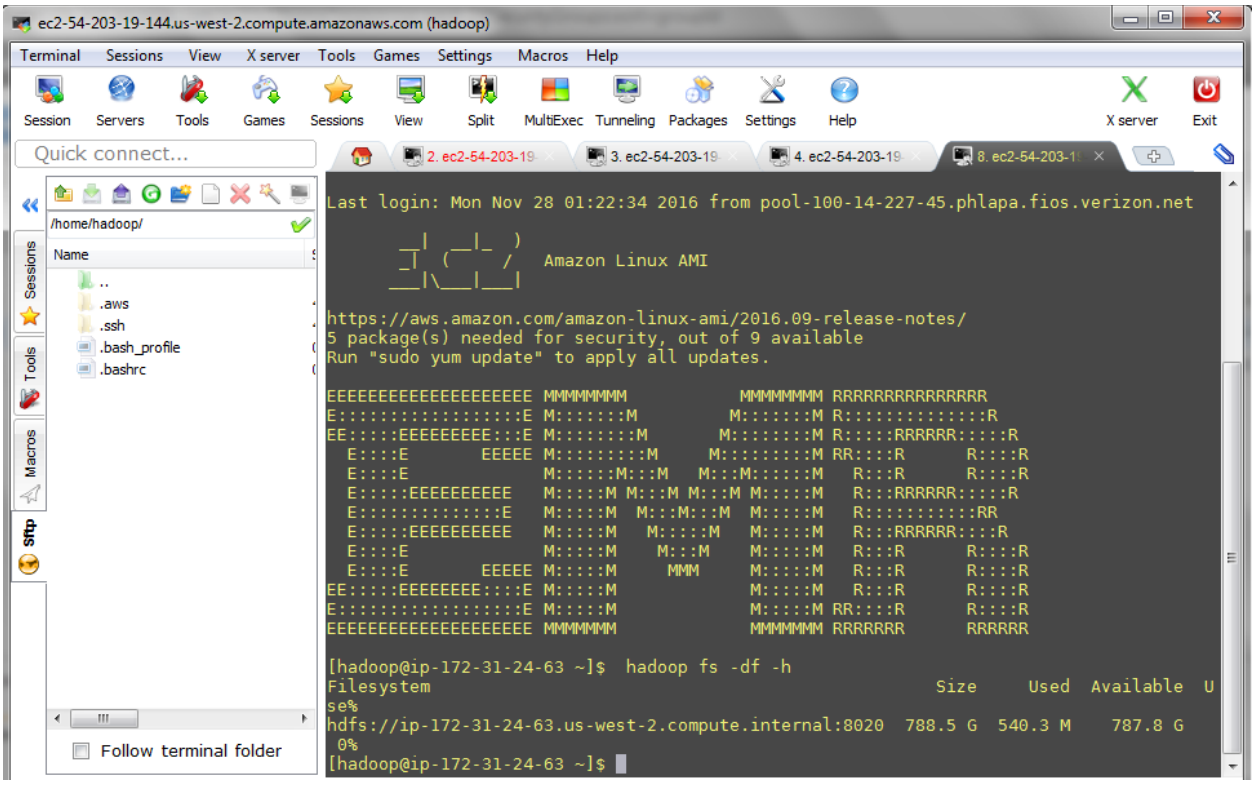
Step 4: Wait for the Cluster to Auto-provision

AWS will auto-provision the cluster. The nodes will switch to “ready” state when done



Step 5: Grab the SSH Key and SSH Into The Head Node

The hdfs binaries should be in the PATH, hdfs fs -df -h will output the total allocated HDFS storage


ec2-54-203-19-144.us-west-2.compute.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. ec2-54-203-19
3. ec2-54-203-19
4. ec2-54-203-19
8. ec2-54-203-1
/home/hadoop/
Name
..
.aws
.ssh
.bash_profile
.bashrc
Sessions
Tools
Macros
sftp
Follow terminal folder
Last login: Mon Nov 28 01:22:34 2016 from pool-100-14-227-45.phlapa.fios.verizon.net
Amazon Linux AMI
https://aws.amazon.com/amazon-linux-ami/2016.09-release-notes/
5 package(s) needed for security, out of 9 available
Run "sudo yum update" to apply all updates.
[hadoop@ip-172-31-24-63 ~]$ hadoop fs -df -h
Filesystem Size Used Available U
se%
hdfs://ip-172-31-24-63.us-west-2.compute.internal:8020 788.5 G 540.3 M 787.8 G
0%
[hadoop@ip-172-31-24-63 ~]$

APPENDIX 2. Data visualization using Shiny and Tableau

Example of a Shiny Live Datastream Dashboard
Source: https://gallery.shinyapps.io/087-crandash

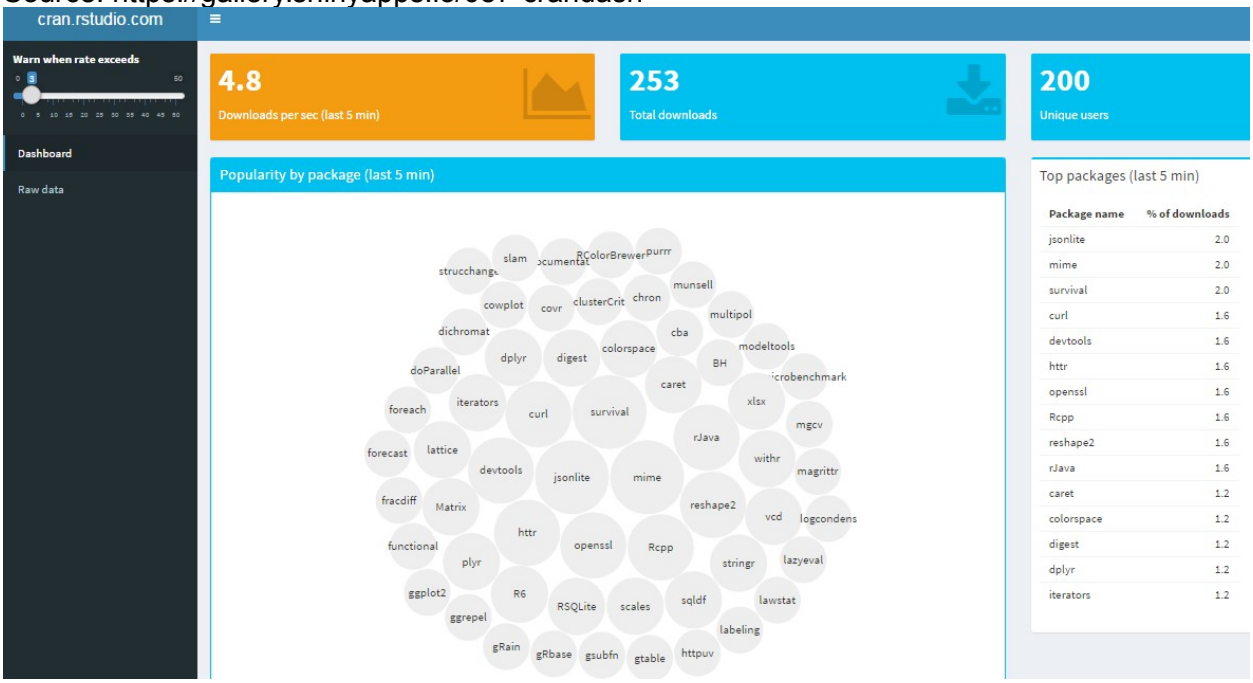


Diagram 6: Example of a Tableau TreeMap Representing Patent Data Related to Cancer Moonshot

Source: https://public.tableau.com/profile/uspto#!