How Do You Consume A Yotta Data? One Byte At A Time!

Dave Kowolenko
Vice President, North East Division Engineering
Comcast Cable Corporation

Kalpa Subramanian
Director, National Engineering & Technical Operations
Comcast Cable Corporation

Contents

What is Big Data?

There isn't an industry conference today where the topic of big data isn't being discussed. So what is the entire buzz about? A search of Wiki describes Big Data as *a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."*[1]
Big data is also comprised of structured and unstructured data elements. Structured data is somewhat predictive and fits well within a pre-defined data structure/model, whereas unstructured data *refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.*[2]

So it is not surprising that the cable industry has arrived at a crossroads where there is more information than classical tools can handle today. Regardless of the industry, accurately interpreting data in a timely fashion is critical to success. In a recent, *EMC-sponsored IDC Digital Universe study, "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"— which found that despite the unprecedented expansion of the digital universe due to the massive amounts of data being generated daily by people and machines, IDC estimates that only 0.5% of the world's data is being analyzed* [3]*.* The article goes on to say that *Machine-generated data is a key driver in the growth of the world's data – which is projected to increase 15x by 2020* [4]*.* The cable industry is a major player in this space by enabling its customers the ability to generate and exchange information in many different forms- video, voice, and data. At the same time, the industry continues to develop and implement a network that provides a vast amount of performance and usage information. As a result, more and more intelligent devices are deployed in the network, from the core to the very edge of the network. These intelligent devices have the ability to provide a vast array of information for a cable provider. It is the challenge of the cable industry to find a way to harness the potential of this information.

Why is it important?

Cable providers or MSO's (Multiple-system operator) are basically being overrun with data at every turn. Yet, at the core, the problem is still the same- increase customer satisfaction, increase customer retention, reduce cost, increase revenue, and innovate! Every functional group within an organization strives for timely insights on how to increase performance and customer satisfaction. Marketing, Finance, Sales, Call Centers, Engineering, and Headquarters are all intertwined and thirst for this valuable information- sometimes looking for that tidbit of insightful information that differentiates them from the competition.

Companies love their data and users often feel the need to store vast amounts of data regardless of the real use. It is hard to manage the accuracy and integrity of such large amounts of data. Not only does this produce erroneous results, it could result in overloading a data storage unit severely enough to cause sluggishness in the entire ecosystem. In today's data hungry world, some systems are designed to hold all data points across various verticals (regardless of its value) and no iota of data is treated as insignificant. This type of data hoarding has caused a newer phenomenon called "WORN" Write once, read never [19].

Since data hording involves such an enormous amount of data, conventional data analysis methodologies are not capable of providing comprehensive or holistic interpretation of the information. In an example from the medical community, numerous studies showed that women who were taking hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD). This conclusion led doctors to propose that HRT was protective against CHD. Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups, with better-than-average diet and exercise regimens. While this conclusion was derived from the information collected, the misconstruction was a result of a narrow view of large data analysis which led to an incoherent correlation. It was only after the data was reviewed by a different end user did the results of the study change [5].

The above case illustrates a situation where certain data volumes were used to correlate against each other. However, in this situation, the data association is characterized by the lack of cross correlation. The above fallacy has the following general structure:

Various data points about an element X are analyzed to determine which data points were correlated against each other. Data points which were identified as being regularly connected because of the number of times correlations was observed inferring that A caused B. However if a third common cause was omitted because of the volume of data, then a wrong conclusion would be derived. Such issues can be prevented through the use of tools and processes that enable the holistic view of the collected data, testing the hypothesis, and by cross connection [6].

The above medical example highlights how data analysis is often tailored towards the stakeholders themselves out of necessity and could fail to associate common trends with the information; resulting in potential lost opportunities for MSO's. Coupled with the complexity and increasing amount of choices that customers have in the market place today, time is of the essence for the industry's need to address this issue now! Companies are starting to rally their organizations to harness the plethora of data in today's global networked marketplace. A report from McKinsey and Company sums up the potential for solving the big data problem:

There are five broad ways in which using big data can create value. First, big data can unlock significant value by making information transparent and usable at much higher frequency. Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore expose variability and boost performance. Leading companies are using data collection and analysis to conduct controlled experiments to make better management decisions; others are using data for basic low-frequency forecasting to high-frequency nowcasting to adjust their business levers just in time. Third, big data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Fourth, sophisticated analytics can substantially improve decision-making. Finally, big data can be used to improve the development of the next generation of products and services [7].

What are we trying to solve?

While some have tried, I think all of us realize that you cannot boil the ocean. Big data is a big opportunity and it can't be solved ALL at once. The challenge is trying to take a vast amount of disparate data and distill it down into small easy to use actionable/insightful pieces that people can work with. It shouldn't be a surprise that the cable network of today continues to evolve at rapid pace. The number of products and services that are offered to customers and the way they access them is growing at an unprecedented rate. Regardless of the product, understanding the customers experience is going to be a real challenge for cable providers. But it is not only the number of products, it is also the number devices that customers can use to access their content. In regards to service assurance, whether it is through a physical or wireless connection, a cable provider must be able to isolate the issue in order to service the customer. The industry needs to develop service assurance tools that keep pace with the architectural changes in the network and customer's demands. The cable industry investment in IP infrastructure has enabled the transition from a distributed proprietary architecture towards an open standard cloud-based solution(s). These standards developed cloud-based solutions will enable cable providers and manufacturers to quickly implement hardware and software that interoperate with each other. These new services are going to infuse the cable industry with more data to manage. Some of the most recent deployments at Comcast include IP multimedia subsystem (IMS), cloud based set top boxes using Comcast's Reference Development

Kit (RDK) software, Comcast Xfinity Home (home automation and alarming), Metro Ethernet, and hosted IP enabled PBX services. Not only are these architectural changes beneficial to the deployment of new products, but it should streamline the development of service delivery and assurance tools which will be used to support these new network elements. These tools can also be developed on an open standards architecture overseeing a standards based architecture - unlike many of today's proprietary network elements.

In this paper, the focus will be on leveraging a big data solution for service assurance and how additional tools are being developed to help identify issues in the network. With big data comes big problems- problems on how to harness the value and opportunities in a powerful and meaningful way. The paper offers up case studies into the speed and multiplicity of data. The size of data storage and speed required for queries constitutes an interesting challenge with respect to performance. Specifically, as one of the largest service providers in the world, this paper offers a unique perspective into ways in which data is made manageable and actionable in the Comcast network.

The goal of any big data solution is to provide a comprehensive view of the interactions between customer care, network performance, network maintenance, provisioning, installation, repair, and vendor performance. Organizations need to be prepared to operate differently than the way they have in the past. Big data is a relatively new concept in the business world today and many managers are going to have a hard time grasping the concept of what it brings to the organization. Big data should equal big positive changes.

Where to start?

While there is a sense of urgency and a long-term vision is key to any project, it also should be recognized that the entire problem cannot be solved right out of the gate. Big data requires the organization to look at data differently. Instead of focusing on where to get the data for your analysis, start out by identifying the business problem you are trying to solve; prioritize a set of issues and focus on a few.

You need an approach that allows for rapid and flexible approach to solving this problem. Many companies (both large and small) are starting to embrace nimble and agile business processes. *Agile software development is defined by WIKI as a group of software development methods based on iterative and incremental development, where requirements and solutions evolve through collaboration between self-organizing, cross-functional teams. It promotes adaptive planning, evolutionary development and delivery, a time-boxed iterative approach, and encourages rapid and flexible response to change. It is a conceptual framework that promotes foreseen interactions throughout the development cycle* [8]. The business continues to change and new challenges will arise daily so once you start this process you should NEVER stop! So it is very important to embrace a culture of continuous "*adaptive planning, evolutionary*

development and delivery, a time boxed iterative approach, and rapid and flexible response to change!" [8] The cable provider will also need to have the proper safeguards in place so that you don't impact the quality and performance of the services that you provide. Being agile doesn't mean that you lose contact with the rest of the organization and cause negative impacts to the business. As the agile approach suggests, don't waste time on detailed design requirements; spend your precious time collecting "use cases". "Use cases" are real-world business examples/problems that illustrate the challenges of managing large amounts of data. These challenges should illustrate complexity, resources (people, hardware, software), timeliness, accuracy, and the corresponding impact to the business. "Use cases" help prioritize where to start and are also helpful in verifying the results from the application. Use cases will also help to identify what data sources you are going to need.

Companies have the option of developing the skill sets in-house or leveraging a third-party provider with these competencies. Time, money and resources will define the approach that you take. In the short term, most companies might be better off leveraging third-party resources since they most likely have the inherent skills necessary to implement a solution in a short period time. Keep the agile software development model in mind as you develop your working team. Once you embark on this journey others are going to want to tag along, so be sure to select a small key group of stakeholders from different areas of the organization that can help drive the project. Remember, the goal for the development team is to exhibit an agile behavior-something that many organizations have not experienced today. Challenge the current business rules with this agile mindset. You might be surprised at how responsive people will actually be. Regardless of where these resources come from, they will need to be dedicated to the process. These resources require individuals that specialize in data modeling, data scientists, application developers, application support, data analysts, and report specialists. In addition, resources that have a detailed understanding of the business processes, data sources, network and data security, and network topology will need to be committed to the program.

There are several companies that specialize with big data applications. Each one has their own unique set of capabilities and every cable provider will have to assess the merits of these benefits as relates to their business problems. This paper is by no means endorses anyone company over the other, rather it is basically a framework by which any provider should attack their big data problems.

In figure 1 Guavas, a big data provider, illustrates the framework of a big data solution [9]. They break this framework down into four different areas: collection and infusion, correlation and aggregation, data modeling and caching, and visualization and reporting.

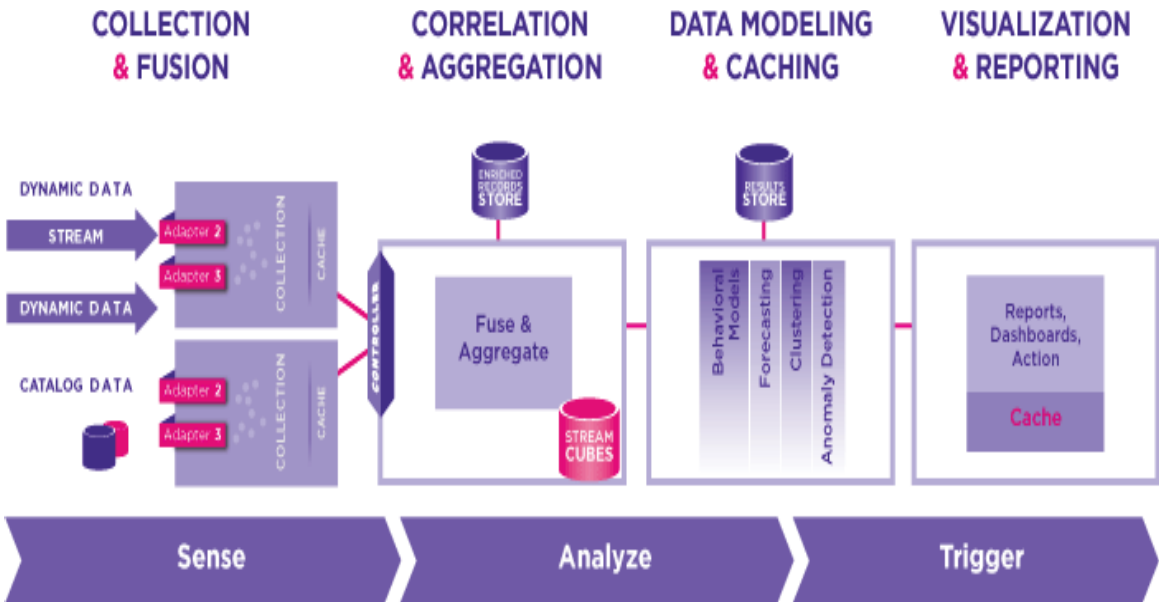


Figure 1: Framework of a big data solution

Collection and fusion: the application must be able to ingest and interpret a vast array of disparate data sources in a timely fashion.

Correlation and aggregation: The application must be able to associate the pertinent data sources accurately and timely.

Data modeling and caching: this is the core of the application. This describes the application's ability to determine anomalies, behaviors, trending, etc.

Visualization and reporting: this is the most critical aspect of the application. The goal with the visualization and reporting is to have the application present information that is simple to understand for the end user.

All four of these elements must work in concert with each other and allow for a rapid and agile deployment approach. Every cable provider must weigh the importance of each one of these different categories when determining their solution.

Show me the data!

Now that you have your team and the use cases have been defined you need to determine where the data should come from. There are several things that characterize data. IBM defines big data in four dimensions: volume, velocity, variety, and veracity. [10]

Volume: the cable industry today produces an enormous amount of data. Data comes from billing systems, our customers, and our networks.

Velocity: describes the timeliness and pace at which data changes. Data could be at rest (static) or in motion (streaming). Understanding and correlating the combination of static and streaming data has to be done in a timely fashion. A perfect example of this is being able to associate call volumes and network events.

Variety: highlights that data can be static, dynamic, structured, and unstructured. Being able to take the different forms of data and correlate them into meaningful insights is at the core of a big data solution.

Veracity: deals with the accuracy and truthfulness of the information being presented. Trying to come up with an accurate view of large data sources is a real challenge for any organization. Data may be replicated throughout the organization and finding the single source of truth may be a challenging activity. Quite often management questions the integrity of reports coming from several different sources. Sometimes similar reports are generated using different parameters; which could confuse the end-user and lead to "paralysis by analysis".

Big Data is often times misunderstood to represent giant volumes only. On the contrary, Big Data really refers to use cases where raw data (several million entries or more) is crunched to provide actual results. The goal of big data being able to split into manageable chunks of data that is extremely fast and immediately usable. Archiving and supersizing data with classical database architectures will cost organizations a significant amount of money. However, using timely data to make and test decisions quickly will help corporations. This concept is BIG SPEED.

The prioritized use cases will help defined the importance of volume, velocity, variety and veracity. Access to data will depend upon the cable providers approach to their implementation. Some big data solutions are externally hosted were others are developed or hosted internally. Hosted solutions by themselves possess a whole litany of issues. Externally hosted solutions require bandwidth, firewall access, scalability, and concerns surrounding the privacy and integrity of the exchanged information. While not insurmountable these issues may delay implementation. Conversely with an internal solution, dedicated hardware needs to be scalable to support the application and if it is hosted internally secure access for the vendor needs to be established. Overall security of this information needs to be considered during the design process; including how information is presented.

Figure 2 shows the simplistic overview of the big data problem focused on service assurance for a cable provider.

Figure 2: Illustration of Big data problem

Being able to distill this information down is a critical aspect of the big data application. Each data source by itself is valuable, but fusing it together to glean new insights could be truly priceless to the cable industry. These golden nuggets are timely pieces of information that a cable provider can leverage in order to prevent an impact to their customers. These are insights that are not easily evident using conventional analytical and/or service assurance tools.

The person behind the curtain

As was mentioned earlier there are several ways to implement a big data solution. Having the right hardware and software are key to any successful implementation. A quick search of the internet lists a vast array of big data architecture solutions. Each one has its own unique set of implementation and performance characteristics. Here is short list of big data solutions:

Massively parallel processing systems (MPP) is a data warehouse appliance consists of an integrated set of servers, storage, operating system(s), DBMS and software specifically pre-installed and pre-optimized for data warehousing (DW) [11].

Hadoop/map reduce - (High-availability distributed object-oriented platform) is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware.

Column oriented database management systems – is a database management system (DBMS) that stores data tables as sections of columns of data rather than as rows of data, like most relational DBMSs. This has advantages for data warehouses, customer relationship management (CRM) systems, and library card catalogs, and other ad-hoc inquiry system where aggregates are computed over large numbers of similar data items [12].

NoSQL Databases - A NoSQL database provides a mechanism for storage and retrieval of data that uses looser consistency models rather than traditional relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. NoSQL databases are often highly optimized key–value stores intended for simple retrieval and appending operations, with the goal being significant performance benefits in terms of latency and throughput [13].

Regardless of the database structure selected, performance tuning will be a specialized area in itself.

Visualization

Visualization is so important that as part of the IEEE's first International Conference Big Data Visualization on October 6th, 2013 in Silicon Valley, CA, the IEEE will be convening a workshop devoted entirely to Big Data Visualization. From the IEEE website, *the ability to make sense and maximize utilization of such vast amounts of data for knowledge discovery and decision-making is crucial to scientific advancement, business success, clinical treatments, cyber and national security, and disaster management. To provide this ability, we need new tools beyond conventional data mining and statistical analysis. Visualization is one such tool and shown to be effective for gleaning insight in big data. However, it is important that this crucial technology will properly address the challenges and match the needs of the users.* In this case a picture is truly worth 1000 words. *This phrase refers to the notion that a complex idea can be conveyed with just a single still image. It also aptly characterizes one of the main goals of visualization, namely making it possible to absorb large amounts of data quickly*[14].

You just can't have a pretty webpage of useless information- it must *properly address the challenges and match the needs of the users.* [15] Using the agile development process, take the time to mock up the visualization layer. Bring in as many end users as possible to validate the information and usefulness of the material being presented.

The visualization picture needs to be relevant to the appropriate end user. It must be intuitive in identifying the anomaly and easy to navigate and possess the capability for the end user to drill down and export information. Figure 3 from SAS is an example of big data visualization [16]

Figure 3: Data Visualization

Most of us would agree that most service providers have a plethora of dashboards and tools for which they operate the business today. Just to create another dashboard is not going to solve the problem. You need to be able to tailor the application to the end-users themselves. Most end-users are not going to be concerned about all of the data, but only that portion which they are responsible for. A typical cable provider will have their network operations center at the top of the service assurance "pyramid" which oversees the performance of the network. With implementation of a big data solution comes the pyramid approach for data presentation. Figure 4 illustrates this concept.

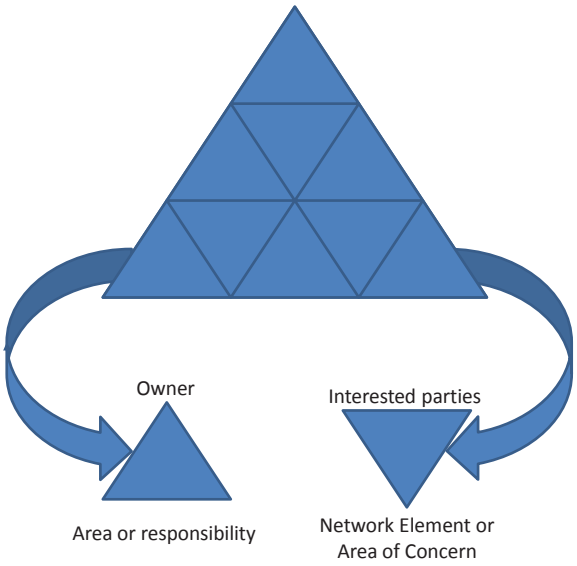


Figure 4: Service Assurance Pyramid

Within any pyramid you can draw an infinite number of smaller pyramids. The orientation of those pyramids describes the end-users. The classic looking pyramid (the point at the top) represents one owner overseeing their area of responsibility. The inverted pyramid shows a number of interested parties (users) focused on a network element or an area of concern. Big data must equal big opportunities. But one organization by themselves cannot fix all of the problems in the network. Real success will be based upon getting the entire organization engaged with using the product. The beauty with the pyramid approach is that you can start with a small number of pyramids and then add more (pyramids/users) as you learn how to leverage the tool within the organization additional pyramids (users) will be created.

Another vital implementation concern within the domain of service assurance, is that it is imperative that insights are detected real-time. The big data application should have the ability to ingest and produce exception-based notification through alerts and/or alarms. Unlike dashboards where end users have to pull the information, the big data application should push anomalies to the end user. These alarms (anomalies) are a call to action for the end user and help to reduce the impact to the business. The big data application should have the ability to automatically determine anomalies, but these threshold should be tunable so that it meets the needs of the end user. Depending on where you are in the service assurance pyramid will dictate on how aggressive the threshold is established. The intent of this approach is to gain engagement throughout the organization- what is important for one end user may not be as important to another.

As part of the implementation, there should be a significant focus on being able to export information for use in classical analytical tools. (You really don't want to have information that is locked within a proprietary solution.) This is important as it may not make sense to spend the time and effort to have the big data application generate reports for ad hoc requests. Therefore there may be times were exporting the data is key for further data analysis activities.

Case study:

The following case study demonstrates an implementation of Big Data where velocity or Big Speed was the primary requirement.

Preface:

As one of the largest MSOs in the US, Comcast proactively tracks the performance of more than 40 million pieces of customer premise equipment (CPE). Multiple key performance indicators (at least 20+ KPIs) are polled from each device. These KPI's are polled at 10 minute intervals via SNMP.

Such a vast amount of polling generates massive quantities of data. This data is made available for use by various business groups for data analytics. For a cable provider to run an efficient service assurance organization, requirements and business rules are written such that service personnel are alerted upon the first sign of an issue. Examples of such instances are:

- During a CMTS maintenance, all CPE associated with the CMTS need to be polled and validated before and after the maintenance for fall out tracking. The KPIs analyzed for such pre and post validation is derived from the data polled by this system.
- When newer product offerings are enabled to customers, such as speed increases, boot files changes are required to provide the enhanced speeds. It is important to track the accuracy of the boot file and ensure appropriate customers received the enhanced changes. This is done by comparing boot files and speeds in the data polled above.
- Polling for status of CPE devices to determine modes such as V4/V6 modes or NCS/IMS platform etc.
-

The following is an elaboration on one such set of analytics involving proactive management of CPE devices. The upcoming paragraphs describe the technological infrastructure required to make the analytics a reality.

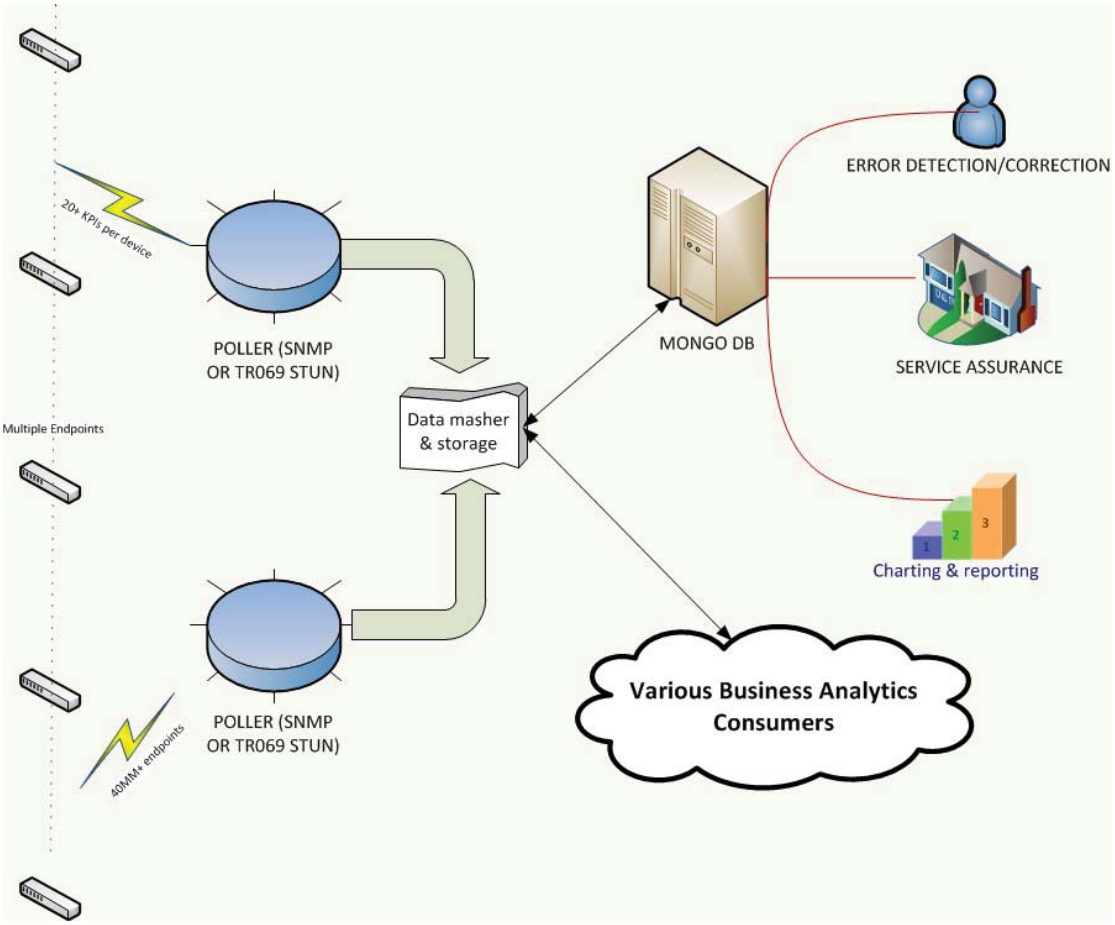


Figure 5: Illustration of the case study environment

Finding the "one":

When designing an infrastructure, several factors were accounted for – Size, speed, row-based versus column based, ACID compliance (described below) etc. As with many projects, one (if not the main) factor was the monetary investment (capital and operational expenditure). Dollar based application/database (DB) decisions inevitably lead into looking for open source options. Choosing an open source option meant that the usual enterprise solutions such as Oracle, Vertica, IBM were out of consideration. Among open source choices, the databases that were not a true RDBMS and not a true MySQL meant Hadoop (or other Apache products), MongoDB, Redis, Riak etc.

The caveat with open source database choices is that while they are moderately easy to do-it-yourself, there is a tradeoff in the process of incorporating more features. For example, to make the ecosystem SQL friendly, add-ons are available, but with each add-on and depending on use, there is a sacrifice to speed and overall performance [17]. No SQL systems often provide weak consistency guarantees and transactions restricted to single data items, even though one can impose full ACID guarantees by adding a supplementary middleware layer. By not providing relational capabilities i.e., strictly

connected data, tables are made nimble and easier to scale. This also helps with the cost associated with data storage by not having to pay the fees associated with relational guarantees[28].

Boiling down to the necessities, the system that Comcast designed needed to be NoSQL, some level of ACID, easy user interface, easy to maintain (by developers, not database administrators) expandable distributed system, and BIG SPEED.
While there are other options that satisfy this set of requirements, our selected choices were reduced to Hbase and associated varieties, MongoDB and MariaDB. Due its relative simplicity, MongoDB was chosen. Having data laid out in the form of documents which can be selectively updated was a huge help. The deciding factor was the ability to simply write JSON scripts to work with the database.

Does ACID compliance matter?

A.C.I.D. stands for Atomicity, Consistency, Isolation and Durability. The specifics requirements to make a system ACID compliant are as follows[18]:

Atomicity – All database modifications must follow an "all or nothing" rule which means that if one part of the transaction fails, the entire transaction fails.

Consistency – Only valid data is written to the database. If, for some reason, a transaction is executed that violates the database's consistency rules, the entire transaction will be rolled back and the database will be restored to a last known good state.

Isolation – Multiple transactions occurring at the same time will not impact each other's execution.

Durability - Transactions committed to the database are never lost. Durability is ensured through the use of database transaction logs that facilitate the restoration of committed transactions [20]

While ACID was considered to be the "rules to live by", systems have systemically moved away from a full ACID compliance. Increasingly more of the cloud based distributed systems are becoming partially ACID compliant, using the cost of full compliance towards speed or availability. In the system that is being designed here at Comcast, with the use of Mongo DB, the system is only partially ACID compliant. However the business requirements imposed on the system do not require a full compliance either.

About MongoDB:

MongoDB is a document-based database system, and as a result, all records, or data, in MongoDB are documents. Instead of the traditional tables in a schema, documents are the default representation of data. Data in MongoDB has a flexible schema [21]. A feature of MongoDB is that it is more cloud friendly through the use of shard. Sharding allows users to partition a collection within a database to distribute across a number of physical servers. When a database's collections become too large for existing storage, users will need to add just one additional machine. Sharding automatically distributes the collection of data to the new server instances or shards [22].

According to developers at Tumblr, which is a company that hosts a blogging system that supporting several billion posts[23], Sharding is the implementation of horizontal partitioning outside of MySQL (at the application level or service level). Each partition is a separate table. They may be located in different database schemas and/or different instances of MySQL. The representation below describes the conceptual function of various shards within a Mongo DBMS.

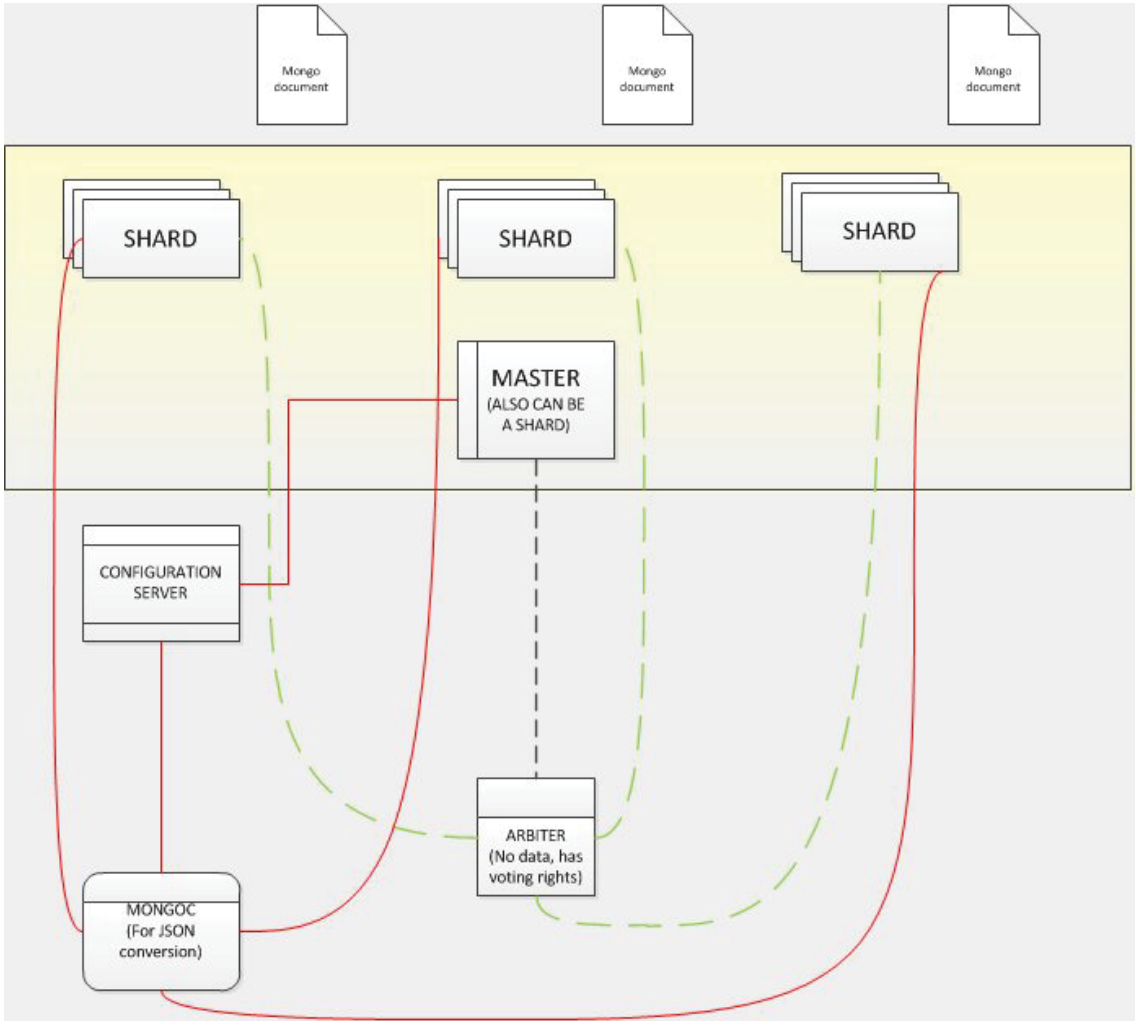


Figure 6: Architecture layout

Achieving data velocity…aka Big Speed:

Quoting The Guardian, Velocity, not Volume is increasingly what determines the hardware and software needs of data-processing organizations[24].

Some of the characteristics of MongoDB that make it agile compared to traditional SQL databases are detailed below. Probably the most important aspect is that there is a cost (time, resources, CPU juice etc.) associated with a structured, fully ACID complaint SQL database. The non-transactional model of MongoDB eliminates that processing cost thereby speeding reads and writes. An aspect to consider is a feature called explicit locking. The code may explicitly lock objects when performing operations. This way, there is serialization of that object, thereby eliminating multiple copy-keeping[25].

Infrastructure

Initial non-scaled infrastructure is depicted below. After the original proof of concept was completed, the ecosystem has since been scaled significantly. The diagram below (Fig 6) is an architectural representation of the described infrastructure. Approximately 18 physical servers are being used in the current proof of concept architecture. There are all 64 bit operating systems running MongoDB. The functions performed by these servers vary including Mongoc, Mongod and Arbiter.

Mongoc lets users compile mongodb queries into Javascript functions that returns true if the query fits the input. Mongod is a shard that holds data. An Arbiter holds no data. However, an arbiter holds a vote to break a tie. While arbiters are called out below as a separate function, they reside in the same physical servers as represented below.

Front-ending the above mentioned architecture is a set of application servers that host a fairly straightforward UI. The UI allows users to pull various business analytics such as volume of modems of a given model, results of a certain SNMP object etc.

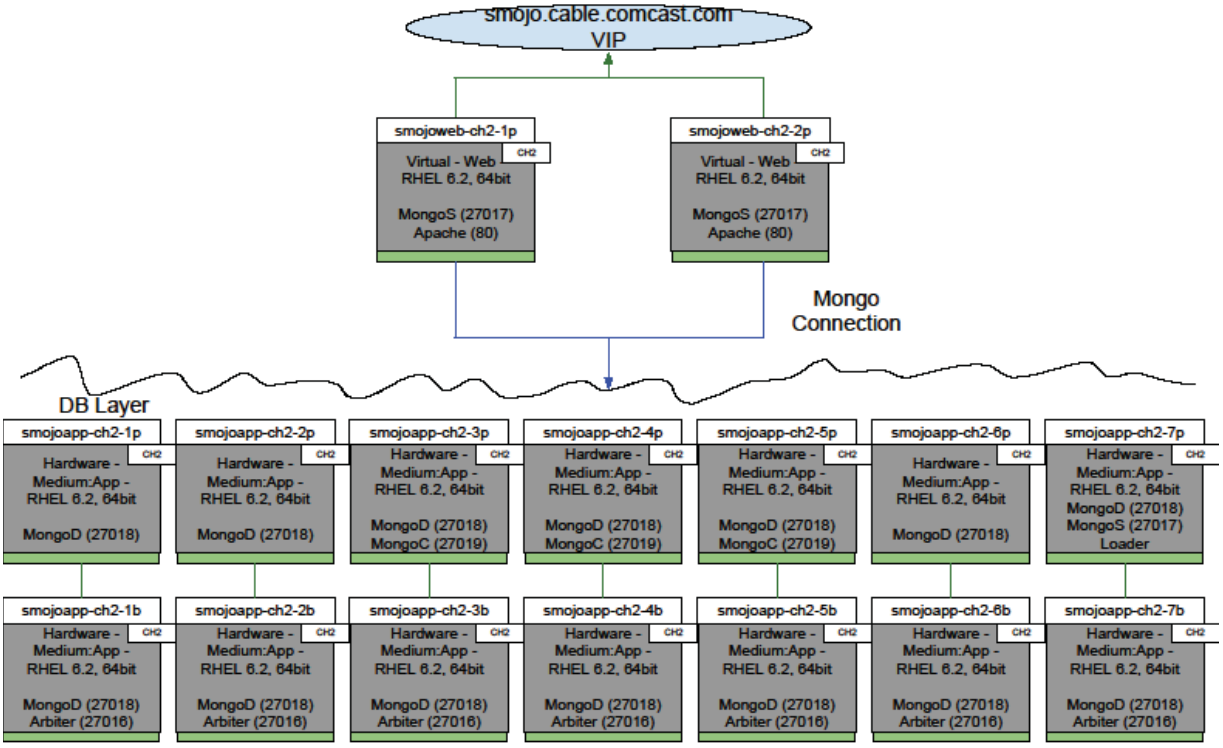


Figure 6: Server layout for case study

Case Study proof of concept:

During proof of concept, queries were run to determine sample speeds [26]. This was compared against a traditional postgres database. The analysis is as follows:

Example 1 - Count of modems with a specific vendor:

Postgresql

- select count(*) from cm_load_1 where vendor
- Return time = 11.4s

MongoDB

- count.json?q_vendor=regex:ARRIS
- Return time = 4.4s

Example 2 – Single modem query for all KPIs collected:

Postgresql

- select * from cm_load_1 where cm_mac='00:15:d1:ba:a5:59'
- Return time = 9.2s

MongoDB

- query.json?q_mac=00:15:d1:ba:a5:59
- Return time = 1s

Example 3 – Complete list of all modems in a CMTS
Postgresql

- select * from cm_load_1 where cmts='cdn16.littleton.co.denver.comcast.net
- Return time = 10.7s

MongoDB

- query.json?q_cmts=cdn16.littleton.co.denver.comcast.net
- Return time = 4.6s

Business rules

While several statistics based approaches are available for analyzing and drawing conclusions from big data, some of the common ones are ensemble learning, predictive algorithms, A/B/N testing, neural networking, pattern recognition, time series analysis etc.[27]

Contextual business analysis

In some of the examples from the Comcast network below, data collected is used in an inferential algorithmic basis. Inferential algorithm in simple terms assumes the following logic:

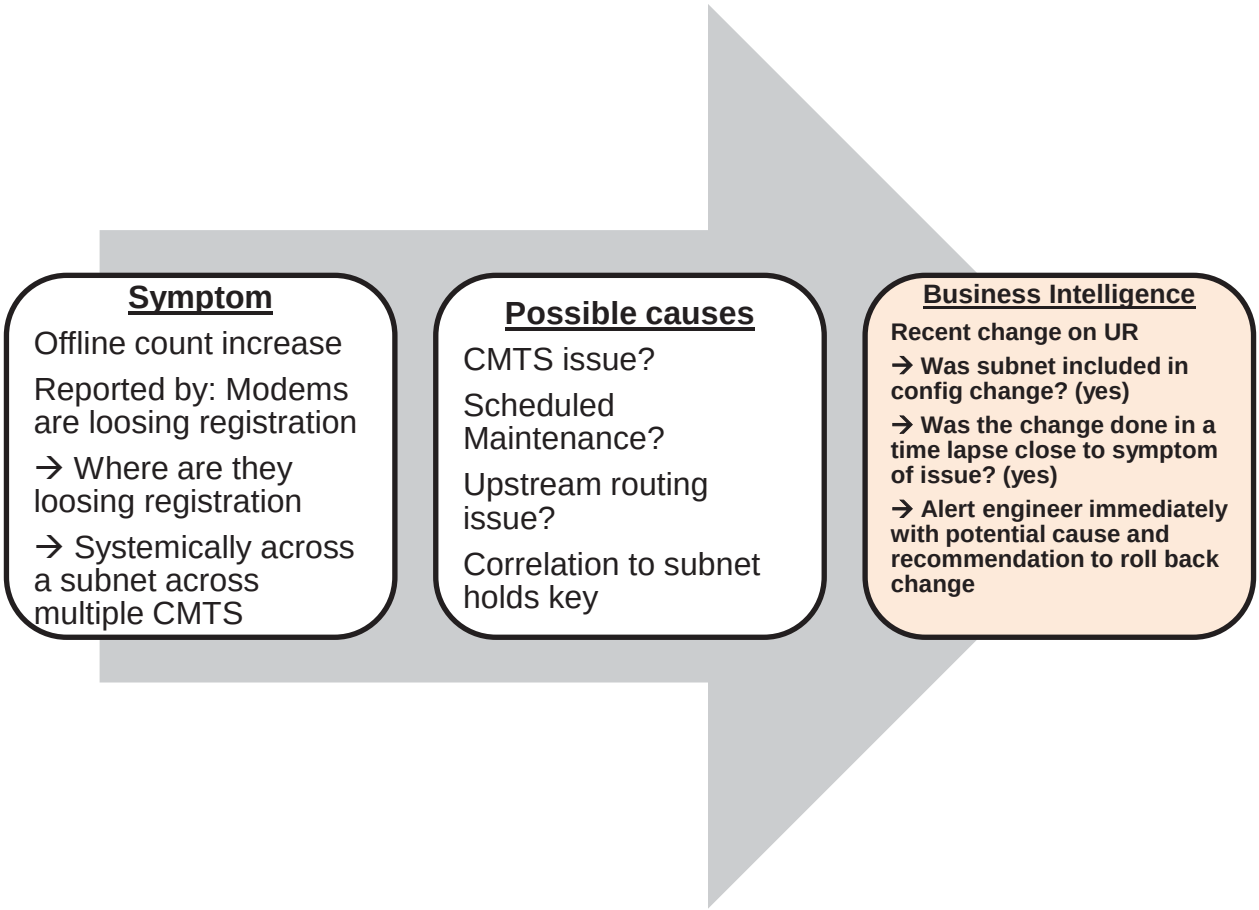
Tasks A and B were performed. It has been determined that an event C is occurring. C could be the results of A or B. Task B is unrelated, and therefore, C is being caused by A. Although the above is an extremely straightforward analysis, it is expected that the user has a large enough sample set
For example: Scheduled maintenance on the billing system was scheduled for date x. Upgrades of the router are also scheduled at the same time. After completion of both tasks, it is determined that subscribers are having difficulty signing up for new features using the customer portal; however the portal is accessible. Therefore, the scheduled maintenance change on the billing could be a potential factor and would need to be verified and ruled out.

Some instances where we use this business intelligence are laid out below:

- Volume of customer calls is collected for each element in the voice network. This volume is charted and trended over several months (learning phase). Business

intelligence base lines the expected volume of calls for each day of week and time of day. Alerts are programmed in such a way that in an event the volume of calls exceeds (or underperforms) threshold alerts are generated. In many cases, this signifies a rogue caller, or a looping situation which needs to be addressed immediately. In some instances, proactive load balancing techniques have to be implemented so as to not overload the network due to increased calling/loading. Such strategies have been utilized during high calling periods such as voting windows after reality shows (American idol, The Voice etc.)

- Loss of dial tone: One of the metrics collected about endpoints is an eMTA or eDVA registration state. In Comcast's IP Multimedia Subsystem network (IMS), packet cable registration status is tracked. Upon identification that a set of eDVAs has lost registration, the logic initiates a series of 'detect and correct' actions as laid out below.



Conclusion

A recent article in the Wall Street Journal illustrated several business success stories. One story in particular highlights the value of embracing big data to solve business problems.

United Parcel Service Inc. has long relied on data to improve its operations. In 2009, it began installing sensors in its delivery vehicles that can, among other things, capture the truck's speed and location, the number of times it's placed in reverse and whether the driver's seat belt is buckled. Much of the information is uploaded at the end of the day to a UPS data center and analyzed overnight.
By combining GPS information and data from fuel-efficiency sensors installed on more than 46,000 vehicles, UPS in 2011 reduced fuel consumption by 8.4 million gallons and cut 85 million miles off its routes[30]

Big data is pervasive and is here to stay. The deployment of new systems and tools will continue to add to the big data challenge. The approach for solving this problem is as vast as the big data problem itself. Identify those use cases or business problems you are trying to solve. These use cases will help you in determining which big data approach would work best for you. Start small, get the right people engaged and be prepared to embrace a cultural change to an agile work environment. So how do you consume a Yotta[29] of data? Take it one byte at a time!

Abbreviations and Acronyms

ACID - Atomicity, Consistency, Isolation and Durability
CHD- Coronary Heart Disease
CMTS - cable modem termination system
CPE- Customer Premise equipment
CRM- Customer relationship management
DBMS – database management systems
DW- Data warehousing
eMTA- Embedded Multimedia Terminal Adapter
eDVA – Embedded Digital Voice Adapter
GIGO – Garbage in Garbage Out
Hadoop-High availability distributed object-oriented platform
HRT – Hormone Replacement Therapy
IMS- IP Multimedia Subsystem
IP – Internet protocol
JSON - JavaScript Object Notation
KPI – Key Performance indicators
MPP- Massively parallel processing systems
MSO – Multiple-System Operator
NCS- Network –based call signaling
PBX – Private Branch Exchange (telephone system)
RDK- Comcast's (Reference Development Kit)
RDBMS- Relational database management systems
SNMP - Simple Network Management Protocol
SQL – structured query language
WORN – Write once, read never

Bibliography

[1], [2] Various authors, Wikipedia. Big Data. Retrieved July 24, 2013, from http://en.wikipedia.org/w/index.php?title=Big_data&oldid=565576788

[3], [4] Weinzierl, Hadley; EMC Corporation. New Digital Universe Study Reveals Big Data Gap: Less Than 1% of World's Data is Analyzed; Less Than 20% is Protected. Retrieved July 24, 2013 from http://www.emc.com/about/news/press/2012/20121211-01.htm

[5] Mayo Clinic staff. Oct 2012. Hormone replacement therapy and your heart. Retrieved from http://www.mayoclinic.com/health/hormone-replacement-therapy/WO00131

[6] Various authors, Wikipedia. Correlation does not imply causation. Retrieved July 24th 2013 from https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

[7], [27] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. May 2011. Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[8], [15] Various authors, Wikipedia. Agile Development. Retrieved July 24, 2013, from http://en.wikipedia.org/wiki/Agile_development

[9] Guavas Inc. Retrieved July 24th 2013 from http://www.guavus-new.com/solutions/platform/

[10] IBM corp, Hurwitz & Associates, Fern Halper. January 2012. Big Data at the speed of business. Retrieved from http://www-01.ibm.com/software/data/bigdata/

[11] Various authors, Wikipedia. Massively Parallel. Retrieved July 24th 2013 from http://en.wikipedia.org/wiki/Massively_parallel_(computing)

[12] Various authors, Wikipedia. Column Oriented DBMS. Retrieved July 24th 2013 from http://en.wikipedia.org/wiki/Column-oriented_DBMS

[13] Various authors, Wikipedia. NoSQL. Retrieved July 24th 2013 from https://en.wikipedia.org/wiki/NoSQL

[14] IEEE, IEEE 2nd International Congress on Big Data. Retrieved July 24th 2013 from http://www.ieeebigdata.org/2013/

[16] SAS corporation. Data Visualization. Retrieved July 24th 2013 from http://www.sas.com/data-visualization/overview.html

[17] Zhou Wei, Guillaume Pierre, Chi-Hung Chi. 2011. CloudTPS: Scalable Transactions for Web Applications in the Cloud. Retrieved from http://www.globule.org/publi/CSTWAC_tsc2011.pdf

[18] OpenBase corporation. Choosing the Right Database: OpenBase SQL. Retrieved July 24th 2013 from http://www.openbase.com/resources/ChoosingTheRtDB-v2.pdf

[19], [26] Schauer, Paul. 2013. CUDA Data Store Overview. Published in Comcast Users and Developers Association

[20] AJ, Marc Intustin. 2010. Is there any NoSQL that is ACID compliant? Retrieved from http://stackoverflow.com/questions/2608103/is-there-any-nosql-that-is-acid-compliant

[21] 10gen Inc. About MongoDB. Retrieved July 24th 2013 from http://docs.mongodb.org/manual/core/document/

[22] Lancy, Doug. Feb 6th 2001. Application Delivery Strategies. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[23] Elias, Evan. 2011. Massively Sharded MySQL. Retrieved July 24th 2013 from http://tumblr.github.io/assets/2011-11-massively_sharded_mysql.pdf

[24] Burn-Murdoch, John. May 2013. Critics of big data have overlooked the speed factor. Retrieved from http://www.guardian.co.uk/news/datablog/2013/may/20/big-data-critics-overlooked-speed-not-size

[25] M, Mike. March, 2011. Why Is MongoDB So Fast. Retrieved July 24th 2013 from http://stackoverflow.com/questions/5186707/why-is-mongodb-so-fast

[28] Mims, Christopher. May 2013. Most data isn't "big," and businesses are wasting money pretending it is. Retrieved from http://qz.com/81661/most-data-isnt-big-and-businesses-are-wasting-money-pretending-it-is/

[29] Various authors, Wikipedia. Yottabyte. Retrieved July 24th 2013 from http://en.wikipedia.org/wiki/Yottabyte

[30] Rosenbush, Steve; Totty, Michael. March 2013. How Big Data Is Changing the Whole Equation for Business. The Wall Street Journal. Retrieved July 24th 2013 from http://online.wsj.com/article/SB10001424127887324178904578340071261396666.html