Journal of Computer Science & Computational Mathematics, Volume 7, Issue 3, September 2017
DOI: 10.20967/jcscm.2017.03.002

Adopting Big Data Analytics Strategy in Telecommunication Industry

Fauzy Che Yayah*, Khairil Imran Ghauth, Choo-Yee Ting

Faculty of Computing & Informatics, Multimedia University,
Cyberjaya, Malaysia,
*Corresponding author email: akunyer@gmail.com

***Abstract*:** Nowadays, adopting big data is a reality. Telecommunication company or telco must find the right solution to store all information available across the organization to maximize revenue using the analytics. The solution must be able to harness the large volume, variety, and velocity of the data available. One of the challenging actions is how to perform decision making and analysis in real-time. Some of the operational decisions may not comply with the corporation policy which makes it hard to keep up with the modern evolving business environment. Telco needs a platform to improve the business process and sustainable and profitable growth. The significant impacts should involve improvements of the customer experience and more reliable network quality, thereby reducing the customer churn rate. Big data and machine learning represent today's trends for the analytics. With big data analytics, the service provider can utilize the full potential of their data set by correlating, processing, and deciphering the hidden information from it. The conventional machine learning tools without big data are becoming inadequate as the trends shift towards distributed and real-time processing. The service provider needs the solution big data-driven which supports them to achieve timely manner and more accurate insights via the predictive analytics, text mining, and optimization. This paper also explains the characteristics of big data, and several uses of case implementing machine learning inside the big data platform related to telco operation such as mobile fraud detection. A well-known big data processing framework such as Hadoop indicated that there is an integration with machine learning tools such as Mahout, H2O.ai, R-Hadoop components, and KNIME. The advantages of these tools are evaluated based on their scalability, ease of use and extensibility features.

***Keywords*:** big data, machine learning, telco data set, pain points, use cases.

1. Introduction

The advent usage of smartphones, internet broadband, and peer-to-peer traffic, social media chatter and IoT (Internet of Things) contribute to the data volume and bandwidth consumption. The internet usage pattern and demand has changed, the explosion of the information is mainly driven by consumers who necessitates for low latency and interactive services. Telcos have been exposed to their bits of data based on their large subscriber connected to networks every day. By extending their voice of customer service into the business, they are now capturing more and more data. This activity leads to big data characteristics from machines logs, internet activities and usages of customer mobiles phones.

However, the understanding of how big data technologies analytics adopted inside the telecom industry still can be ambiguous. This article will fill the gaps in the implementation of big data technologies and some sample use cases inside the telecom industries.

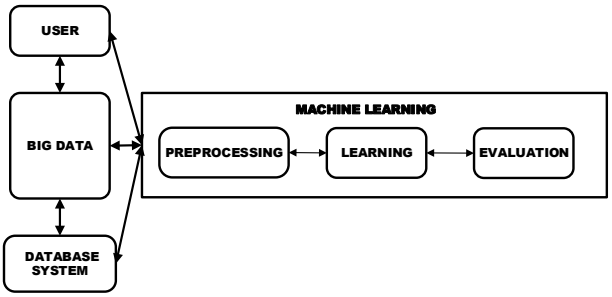

Figure 1. Big Data and Machine Learning Integration.

Machine learning is important by supporting data scientist to solve the computation problem [1-7] efficiently. By combining big data and the machine learning techniques, it is potential to exploit what is "hidden" inside big data. The analytics results are machine driven and operate at machine scale. The more data is fed into the machine learning system, the more it produces a better quality of analytics. Unlike traditional analytics, machine learning thrives and work better on growing data set. Decision tree, association rule, neural networks, deep learning, support vector machine, clustering, Bayesian networks, prediction, and classification are the examples of machine learning approaches [8]. In most of the telco use case, prediction and classification approaches are widely used especially to solve problems related to the customer, networks, and services. There are several definitions of the big data, but based on the literature, the best approach defining big data is by characterizing it. Three key features (3V's) of big data [4] are as follows:

- **Volume**
  Volume is the determined by the size of the data. The large data volume necessitates redesigning of data processing pipelines so that conventional analytical tools may be used. The large data volume also had ramifications on the storage and processing of such data.
- **Velocity**
  Velocity is described as the rate of data flow of fresh data continuously in a real-time from various sources such as sensors and machine logs. As the arrival speed of data increases, it imposes new requirements on storage and processing frameworks.
- **Variety**

Variety is defined as various data format which is coming from the different data sources. For example unstructured, structured, semi-structured, audio and video.

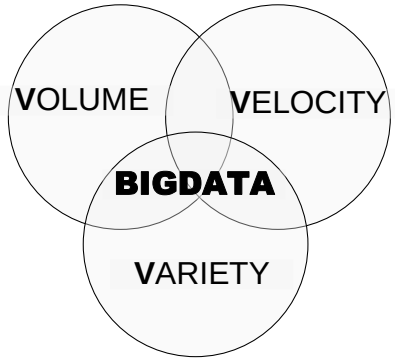

Figure 2. The basic 3V's of Big Data

1.1 Conventional Approach and Hadoop Approach.

Most typically, in service provider customer-centric use case, sample and model are used to classify a particular type of customer based on specific type of their profiles. The difficulty with this approach, although it works, the granularity is not at the level of an individual. Classification based on the segment is good, but to make a decision based on the individual is better. To implement this, working with larger data set is conventionally possible via traditional approach. In some cases, only a few percentage of the data set actually can be used for the data modeling. The limitation is due to cost of the proprietary machine storage and computation [9]. Most of the traditional database resides on a single machine.

For example, the conventional database or relational database management system (RDBMS) finds it challenging to handle such a huge data volume [10]. It needs to be upgraded by adding more CPU and more memory vertically. "Big data" is generated at high velocity and mostly is an "unstructured" format. The unstructured format typically is a text-heavy but may contain other information such as number and dates. However, RDBMS is not designed for this purpose. It only has single query engine which only can implement on its database schema. It works best with structured data sets such as financial data.

Hadoop is an open-source framework for distributed data processing and storage using the MapReduce [10] for big data architecture. MapReduce is the programming model of Hadoop for large-scale data processing. Hadoop is designed working with various types of engines for particular data processing model. For example, for batch processing model, MapReduce is suitable for handling huge data sets when processing time is not a factor. For interactive queries which require a faster result, Hive on Spark is the suitable engine for this task. Real-time data set such as machine logs, live streaming data [11] from the Apache Kafka, require Spark core extension known as Spark Streaming [12]. Spark Streaming enables scalable, fault tolerant, and high throughput capability of the live data stream. In conclusion, the service provider should take into consideration moving toward the Hadoop base approach. The selection for the right data processing engine is dynamic and beneficial for many use cases inside the telecommunication industry.

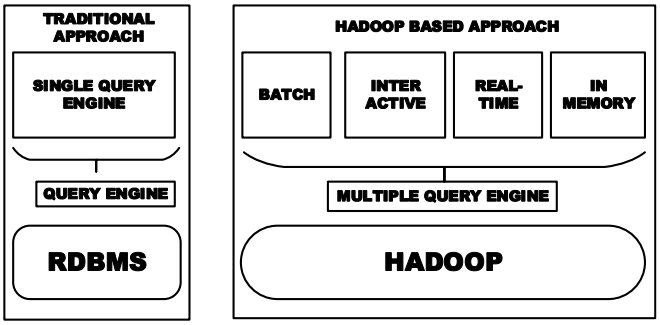

Figure 3. Big data and Machine Learning

1.2 Big data Analytics and Legacy Analytics

Customer-centricity rely on three primary pillars: efficiency, insight, and performance. We are now living in a world that many customers interac via the social media and chat about their internet experiences and related issues online. The telco needs to proactively collect and analyze data for further actionable taken on customer retention and offer more attractive services. However, when dealing with the social media, they are related with unstructured data. The data must be processed before it can be applied to any business insight. Big data analytics platform tools have the capability to correlate for any trends and associations. It is a very cost-effective approach without any data conversion into a structured design similar to the conventional data warehouse which only suitably executes the legacy analytics that is sluggish and has limitation [13]. The table below shows the comparison:-

Table 1. Legacy Analytics and Bigdata Analytics

	Legacy Analytics	**Big Data Analytics**
Storage Cost	High	Low
Analytics	Offline	Real-Time
Utilizing Hadoop	No	Yes
Data Loading Speed	Low	High
Data Loading Time	Long	Faster
Data Discovery	Minimal	Critical
Data Variety	Structured	Structured , Unstructured
Volume	Terabyte+	Petabytes+
Complex Query Time	Hours/Days	Minutes
Data Compression	Not Matured	Avg.50% More
Support Cost	High	Low

There is a requirement for a solution that can consolidate customer related data sets with the network information which helps enhancing the overall customer experience. Big data analytics and machine learning can secure the return on investment (ROI) [14] for the telco. Big data is not a replacement for the conventional analytics infrastructure, but it acts to fill the "gaps" in between for more relevant information. This evolution creates a symbiotic relationship among the current silos information across the organization. The list of the main benefits of the big data approaches will

impact the reduction of the administrative cost of the IT operation, increase the data loading speed and reduce queries execution time of concurrent users.

1.3 Business Value of Big Data Analytics.

Value-added service such as over-the-top (OTT) player to the current service operator is essential. The number of OTT application increases every day due to the content of the mobile application inside the smartphones. An example of a mobile application such as Skype for mobile is one of the customer preferences for a cheaper call [15] and short message service (SMS) via the wireless mobile telecommunication technology (example: 3G, and 4G networks). Telco needs to ensure the OTT communication via their network that are stable for improving the customer experience based on the key business objective. This action will lead towards increasing the revenue, reducing the churn and operation expenditure (OPEX).

Simplifying the business operation is a very significant effort for telcos. Telco always strives to grow more collaborative and to break their data silos [16] by adopting the best practices and methodology to become more competitive. They are looking for the unique value to differentiate themselves with other competitors. Handling the customer experience approach must be efficient to address challenges arise in customer problems.

Telco operator believes that adopting big data technology with machine learning capability will play a critical role to achieve the business objectives. The hidden information within the large telco data set offers such greater value. Such information cannot be explained through the conventional data analysis. It requires a combination of distributed data processing, pattern matching, and additional machine learning methods. Large data sizes and complexity have given rise to new challenges and avenues for research.

1.4 Big data Analytics: Telco Use Cases

A European telecommunication group has utilized big data platform [17] for scalable on-demand analytics. Several system databases have been consolidated into a single location. The enormous amount of data stored in a granular fashion makes it very cost effective. The ability to generate reports directly from big data platform is the most important key factor. The business user accepted it with confidence.

One of the global mobile communication service providers in Europe wanted to understand the customer travel pattern to support real-time promotion, up selling and advertising. Besides that, they wanted to improve the quality of service of their network operation [18]. They were using big data platform with the machine learning capability for the solution. The historical data is stored in raw format. The system ingested 10TB per hour in real-time streaming mode. As for the results, customers realized the service improvement quality as it was capable of informing in real-time experience for any dropped call.

A large telecommunication company supported by Deloitte implements a big data platform [19] to collect, store and analyze billions of customers' transactions to achieve real-time marketing effectiveness. The solution can reduce the data latency [20] from 45 days with the current operation data. With the big data analytics, the company has a better understanding of their customers. Lastly, the company will utilize the social media to understand more about the social sentiment to faster response to any customer issue via customer relationship management (CRM).

Most of the telco do not have fully 360-degree [21] views of their data; they collect essential information such as network performance information, device information, and usage transactions. Most of the data is stored in silos database in a different department. Thus, the data quality can be disputed, and needs a lot of data processing for cleansing correctly for any records duplication and errors. There are also some political barriers for the data set to share across the organization.

Telco needs to have the ability to leverage and exploit their customer information as the key to stay competitive in the market. What is the best strategy for the telecommunication company to monitor and monetize their customer information effectively? Big data demonstrates the capability to centralize all the information for automated data processing and advanced analytics with machine learning. Examples of big data use cases in telco industry can be described in the following sections:-

a) Network Infrastructure

Network capacity planning utilizes big data and machine learning to optimize the rollout of new service coverage [22]. The operation allows the telco to identify the problematic network location which can cause the problems. A large number of related network logs have been collected and will be analyzed for root-cause analysis with the big data platform in real-time. Moreover, for the network upgrades, big data analytics can support the telco that identifies the nodes responsible for network congestion which may impact the customer experience and the churn rate. The analytics results are based on the correlation between the customer and the performance of the network.

In the network maintenance point of view, machine learning helps the telco to maintain the particular equipment proactively driven by the analytics results. It is a cheaper initiative due to some of the equipment located at the remote site, and this action is less disruptive [23] than replacing the equipment that has already failed. For the network performance management use case, more historical records were consolidated into the big data platform. The telco now can learn more granularly the root cause than ever before. The results are beneficial to the network traffic shaping analytics which requires immediate action for any unusual activities inside the network. This problem may hog the network bandwidth and also erode the service quality that impacts the customer experience.

b) Service and Security

Call Details Records (CDR) use case in telco utilizes big data analytics platform to detect the pattern in any dropped calls and low-quality voice [24]. Hundred millions of CDR records ingested into the platform, require a bigger storage. Hadoop File System (HDFS) is capable to store the CDR in compression mode (example, Parquet, and ORC format) to reduce the original file size. The compression format loads faster into the memory. The CDR records are being analyzed

for compliance checking and billing validation for any suspected fraud activity. With large capacity storage, telco now can retain and analyzed the historical CDR data up to three years for continuous service improvement.

Regarding telco contact center operation, the contact agent usually does not have any performance until granular level insight interact with the customer. Thus, they were unable to provide effective call resolution due to time constraint call. By having big data analytics and machine learning capability, it can free up the agents time more by detecting the actual cause of the problem without longer conversation. Some of the calls can be automatically off-loaded to the interactive voice agent for certain low severity cases.

For some cases, the limitation of the contact center agents to diagnose the problem in detail may lead to many unnecessary truck rolls. They need to be informative to decide if the problems happens due to the network quality at the customer endpoints. Sending the truck rolls to the client's premises without understanding the real issue will increase the operation cost. Without big data analytics, it is hard to avoid any false positive situation. Machine learning can help classifying the first diagnostic level before sending the truck roll for further investigation.

Real-time malware threat detection is a use case that is related to the customer device security [25, 26]. Malware attacks can cause network congestion, data leakage, and phishing attacks. Some devices affected by the malware may violate the telco fair usage due to high usage of bandwidth. By utilizing big data analytics platform, the telco can identify the irregularities inside the network. Some remedial action can be taken by informing the affected customer about the service interruption or blocking the device activity temporarily to use the network.

c) Sales and Marketing

Telco's interact with their customers across many channels [26] over time. It is a very challenging task for them to correlate the customer's activities (example: online purchasing pattern, online behavior trending) when the historical interaction records are stored in silos. The problem arises when the volume data increases. It is technically difficult to preprocess this data set inside the conventional relational database. Big data offers enterprise data lake which makes a 360-degree view of customer when historical information is possible to analyze. Predictive analytics with machine learning can discover what next customer may purchase. Voice of the customer's records can be consolidated from the social media website such as Twitter and Facebook.

As for the marketing use case, marketers found that it is hard for them to tailor the needs of each customer with their marketing campaign. Telcos are looking for ways to find information about the customers' interests and behaviors through browsing activities using mobile phones and tablets. By using big data analytics, the telco can pinpoint at the particular customer who may risk of churning. The selection of customer can be part of the marketing campaign to retain them. The telco also found the solutions to segmentize the customer profile. By targeting the "right" customer, it can reduce the subsequent churn for that segment. Thus, the revenue leakages can be minimized.

d) Business

From the telco business perspective use case, it is concluded that new product development [27] is important for providing new value to the customer. One of the business problems is the data that is unavailable (never collected), or it is never transformed into business insights. To overcome this issue, the telco needs to create new pilot projects to identify the pattern of usage from the customer mobile phones and tablets. In every topic regarding the customers' behavior (where, when, how and why) devices are used in collecting the details into the big data platform. As a result, the telco successfully increases the revenue from the analytics been produced.

In other industries such as financial services, big data has been driven entirely by a new business model. Financial trading services now probe a massive amount of market data in real-time to identify opportunities and values from it. In the retail sector, big data support the analysis of customer purchasing behaviors. The stores now can adjust merchandise and stock levels for maximizing the sales and profits. Every industry uses different approaches and is interested in the transformation strategies that utilize the big data analytics capability. The following table shows the area of improvement which utilizes big data:-

Table 2. Big Data Adoption Improvement Areas

	Telco	**Retail**	**Financial Services**	**Energy**
Pain Point Areas	• Customer Relationship and Experience. • Network Infrastructure • Service and Security • Churn Prevention • Marketing Campaign • Fraud Detection • Product Development	• Customer Relationship and experience. • Store Location Optimization. • Supply Chain Optimization.	• Trading Optimization • Risk Analysis • Fraud Detection • Portfolio Analysis	• Smart Grid Anomaly Detection • Power Line Sensors Data Acquisition • New Exploration • Operation Modelling and Optimization

1.5 Big data platform as a shared services.

Hadoop implementation typically begins with the analytics application. As more new applications are created, Hadoop acts as shared services for delivering insights on a different scale. The use of Hadoop can be a complement to existing data system because of its low-cost storage price, data processing power, and scalability capability [28]. With the continued growth number of analytics applications and data, the concept that Hadoop becomes the data lake starts to materialize. Combining data from multiple data sources it helps telco to find the answers from the complex questions previously.

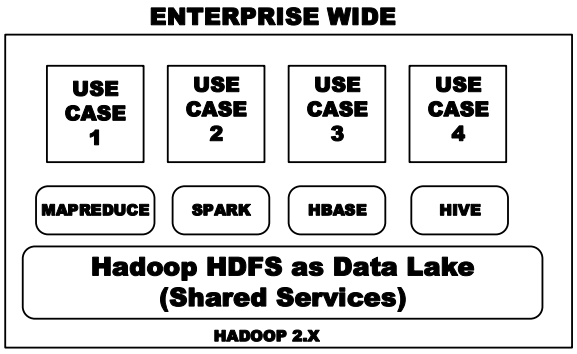

Figure 4. Hadoop as a Data Lake

Hadoop as a Data Lake contains all the data including processed and unprocessed in the same location [28]. It allows users across the business data set to define, enrich, and explore the data set by their preference (batch, interactive, real-time). The shared services are beneficial to the telco operation which are nearly similar to the cloud infrastructure. The speed of provisioning reduces the operation complexity and harnesses the learning curve for the end users. The enforcement of the data security, privacy, and governance inside the data lake are consistent. More data in the same place leads to better insights by running extensive series of analysis. The cost to retain data is lower as the value of the processed data grows exponentially.

2. Telco Data Type

The primary goal of data mining for analytics is to explore the search for patterns and to discover the relationship among the variables. It consists of three stages which are exploration, model building, and deployment. Exploration stage involves cleaning the data set, transforming the data set and "features selection" from the data set to identify the variables the machine learning can use. Model building is the process for selecting the best model from the data training performance. Moreover, the deployment stage is the process applying and executing selected model with the new data. Before all the activities above can be performed, telco needs to identify which data set they need for the analysis. The data acquisition process needs to established and streamlined into central big data storage. The following information describes the approaches required for each type of the data set:-

2.1 Time Series

This type is a group of data with a set of points arranged in time scale. The data point is consistently arranged in repeated measurement over time. Internet of Things (IoT) devices contributes a lot amount of time series data set, and the number increases rapidly every day. It is also important to discover the anomaly, trending and any performance issues from this type of data set.

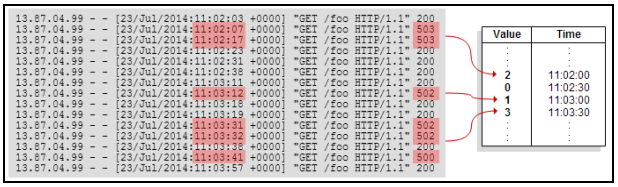

Figure 5. Anomaly detection over time series data set

Examples of such data sets can be traffic utilization of subscribers and network activity. For a data set which has a lot of variations, both clustering and classification techniques are suitable to implement. Support vector machine (SVM) and Random Forest (RF) are the examples of the machine learning algorithm. Other possible tasks that can be performed with the time series data set are the time-series segmentation and motif discovery. Time-series segmentation is the process to divide the discreet data set into smaller pieces segment.

2.2 Streaming

Streaming data can be described as the continuous data creation at high speeds originally from different sources. Examples of streaming data are network activity such as equipment alarms, website logs, Dynamic Host Configuration Protocol (DHCP) or Domain Name Servers (DNS) logs. The data mining techniques for the streaming flow is a computationally powerful process. The conventional machine learning algorithm was designed for the static data set usually does not fit to handle large variables size. This restriction is due to the incoming rate of the data that needs to suit the machine learning algorithms rapidly. Some of the algorithms need to be improved for supporting the data stream. Apache Spark Streaming is one of the suitable Application Programming Interfaces (API) for this job [29]. Apache Spark Streaming is capable of ingesting, processing and analyzing at high velocity of the data.

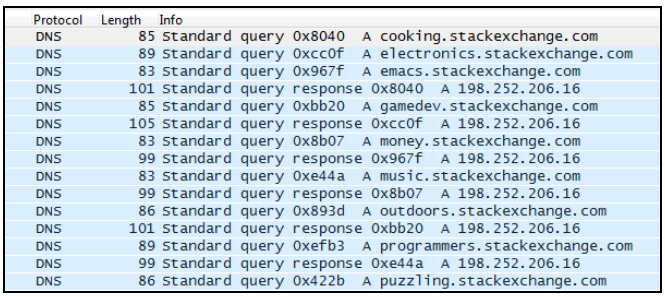

Figure 6. Example of DNS Log

Apache Spark Streaming API discretizes streaming data into smaller parts and generating the output in micro-batches.

Each micro-batch data is represented in Resilient Distributed Data set (RDD) format. RDD also can be accessed by other programming languages by using the API. Another main advantage of Apache Spark Streaming is being capable of balancing the task partition, data structure combiner, and crash recovery during the computation process inside the memory.

2.3 Unstructured

Complex data like unstructured types will not fit into the Relational Database System (RDBMS). It cannot be stored in rows and columns schema. BLOB or Binary Large Object is the only approach for storing unstructured data inside the RDBMS database.

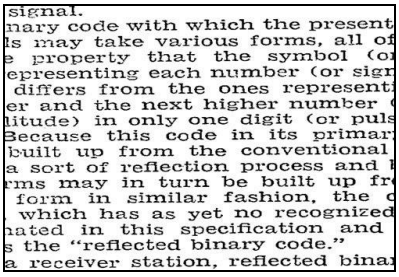
signal.
nary code with which the present
ls may take various forms, all of
e property that the symbol (o
epresenting each number (or sign
differs from the ones represent
er and the next higher number
litude) in only one digit (or puls
Because this code in its primar
built up from the conventional
a sort of reflection process and
rms may in turn be built up fr
form in similar fashion, the
which has as yet no recognized
nated in this specification and
s the "reflected binary code."
a receiver station, reflected binar

Figure 7. The Unstructured Data set

Significant challenges arise from unstructured data mining starts and from the strategy of storing until applying them for decision making [30]. Other related issues are the technique of formulating the unstructured data into a structured format that can fit into the database. The volume of the unstructured data is increasing due to the usage of the social media.

3. Big Data Analytics and Data Mining Tools

Along with the exponential growth of data, it takes a process to extract valuable information from the data and convert it into readable and usable form. This is where data mining comes into the picture. There are lots of tools available for data mining tasks that embedded with artificial intelligence, machine learning and other techniques to extract data. Below is the example of the tools available:-

3.1 Apache Mahout

Apache Mahout is a Java based distributed scalable machine learning and data mining tools. It can work on Hadoop with different machine learning approaches such as clustering, classification, and collaborative filtering. It only utilizes MapReduce framework for distributed computation. List of algorithms supported by Apache Mahout includes items such as Logistic Regression, Naïve Bayes, Hidden Markov Models, K-Means and Support Vector Machine (SVM) [31]. The purpose of this tool is to improve the analytics that applies the big data environment into more significant information discovery.

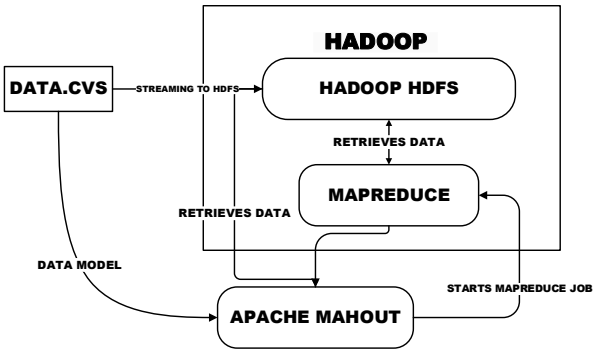

Figure 8. Apache Mahout and Hadoop integration

3.2 Konstanz Information Miner (KNIME)

KNIME is built on top of Eclipse platform for software portability. It has many extensions for data mining such as Bayes , Clustering, Neural Network, Decision Tree, SVM and others. It also provides additional functionality for data filtering, modeling, and also visualization [32]. One of the advantages of KNIME is integration with the R and WEKA library. Any analytics model that is built with R code can also be performed inside KNIME tools. KNIME big data component such as KNIME Spark Tool Integration supports data processing with Apache Spark inside the Hadoop platform.

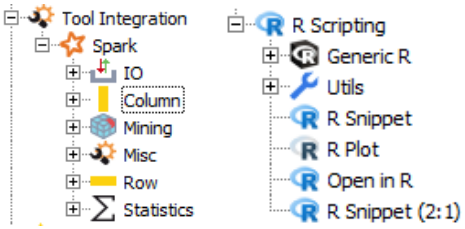

Figure 9. R and SPARK tools integration in KNIME

3.3 R Language

R is a language for the statistical computing and graphics [33]. R-Studio is the IDE for R which operates on the most operating system including Microsoft Windows and Linux. R store all the computing objects inside the memory, and this functionality can become a restriction if the data set is too large.

RHadoop developed by RevolutionR Company, succeeds to unlock the in computing limitation in memory [34]. RHadoop is R package that can integrate with HDFS. RHDFS enables R developer accessing HDFS and does modification inside it. Another package, RMR2, allows the developer to perform custom MapReduce code by using R. The problem of interfacing with Hadoop is the complexity [35] to write the code. Nevertheless, R is not as mature as JAVA programming language. Combination of R and Hadoop can become a perfect solution for data crunching tools for serious big data analytics for business. Using R on Hadoop will contribute scalability advantages on data analytics without limitation [35] on the size of data set. This benefits will let data scientist working on data set that is larger server's memory in parallel computation. Combination of R and Hadoop in big data analytics will serve in returning cost of value due to it fits to run on commodity hardware for future vertical scaling. Both are open source

components which are low cost and avoid the vendor software proprietary lock-in.

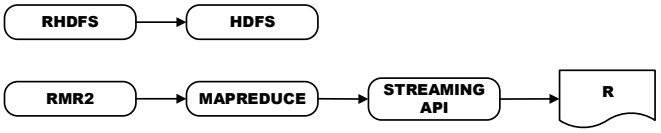

Figure 10. RHDFS and RMR integration

3.4 RapidMiner

RapidMiner provides a powerful user interface that enables the user to create and deliver predictive analytics in an uncluttered approach. This functionality will minimize the usage of manual scripting and reduce the common errors made by the developer [7]. It also supports the data integration, transformation, and machine learning functionality. RapidMiner Radoop is the plugin of RapidMiner that compiles execution code which transposes the process workflow into MapReduce or Spark enabled inside the Hadoop, therefore it unlocks the large data size computation limit.

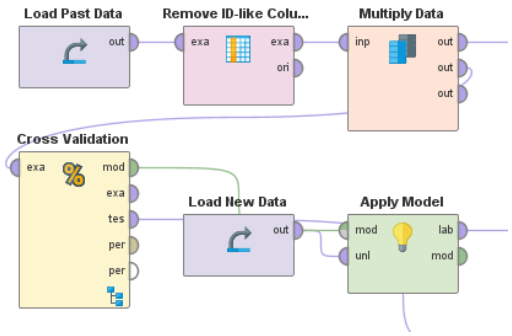

Figure 11. RapidMiner Process Workflow

Table 3. Comparison of Big Data Analytics Tools

Analytics Tools	Hadoop Integration	License	Solutions	Users	Version	Language	Execution Method
R	R Library Installed	GNU GPL 3	Scientific Computation / Data Mining / Machine Learning	Very Large	3.4.1	R , Fortran , C	GUI & Command Line
KNIME	Licensed Version	GNU GPL 3	Data Mining / Machine Learning	Large	3.4	JAVA	GUI
RapidMiner	Licensed Version	Open Source	Data Mining / Machine Learning	Large	7.5	JAVA	GUI
Apache Mahout	Yes	Open Source	Machine Learning	Medium	0.13	JAVA	Command Line

4. Telco Related Data Set Structure.

In this section, the basic structure of telco related data set is defined and discussed as follows:-

4.1 Subscriber Data

Subscriber data in one of the most critical sets in telco. The data set contains various type of information regarding the customer profile. The data set contains regular columns as:

- Subscriber billing information (name, address and
- location)
- Installation date
- Termination date
- Premises details (*address, state and zone*)
- Subscriber package (*internet speed package*)
- On-premise device type (*model and technology*)
- Fiber or Copper Distribution Point (DP)
- Reseller information
- Payment details , and
- Exchange building information (*location*)

Telco's subscriber data is classified as one of the characteristics of big data by its large number of subscribers. By increasing number of subscribers, more information such as network activity, network inventory, network quality, customer buying history, and customer downloads activities can be obtained. A massive amount number of subscriber data set requires preprocessing capabilities using Hadoop for aggregation. Hadoop data acquisition components such as Sqoop is used to stream in [36] for any changes based on data sources Change Data Capture (CDC) insert or modified date. The parameter below is required for Sqoop operation which enables the CDC functionality below:-

Table 4. Sqoop parameter with Change Data Capture (CDC)

Argument	Description
--incremental	Defines the column to be scanned and rows to be imported
--check-column	Defines which rows are modified
--last-value	Last value from the last import session

All the imported data will undergo for preprocessing stage and data cleaning. This activity can be accomplished using Apache Hive and Apache Spark extension. The final aggregated version data set now can be accessed from external Business Intelligence (BI). These tools are required for business report generation.

4.2 Subscriber Trouble Tickets

Trouble ticket system is a software for tracking and troubleshooting [37] all incidents or issues reported by the subscriber for any service disruptions (i.e. telco services). If the problem is identified, the call center will issue a new ticket with a unique reference number for the customer for their future reference. The call center also will enrich more information about the severity level, account number, transcribed conversation in details for the remarks (*Description*) section inside the trouble tickets system.

Table 5. Typical of the trouble ticket data set column

Column Name	Value	Description
Tt_Row_Id	1-2dltgzn	Unique Row ID
Tt_Num	1-abbcc71939	Trouble Ticket No
Tt_Type	Cust.Tr.Ticket	Trouble Ticket Type
Tt_Sub_Type	Proactive	Trouble Ticket
Status	Closed	Subtype
Severity	4-Low	Trouble Ticket Status
Imp_Message	1st Check	Severity Status
Appointment_Flag	Yes	Important Message
Account_Name	John	If Appointment Set
Subscriber_Num	1014110827	Subscriber Name
Account_Num	1-27v1srdd	Subscriber Number
Created_By	X-568789	Subscriber Account
Category	Failure	ID of the TT creator
Description	**< Text >**	Trouble Ticket Category Detail about the ticket
…	…	

'*Description*' is the column contains text information about the complaints and every diagnostics step. This point describes the characteristic of big data characteristics which is 'Velocity' and 'Volume' for this type of telco data set.

4.3 Subscriber Internet Connection Performance

Internet subscriber performance test or the "speed test" contains information such as internet upload and download speed, ping latency, and jitters. Users can measure their internet service performance by the web browser or mobile apps from the telco's official speed test [2] website. The example of the generated columns of the internet performance as follows:

Table 6. Internet Connection Performance Example

Column Name	Value	Description
Date	2017-01-03 19:59:06	Speed test time
Login	yenpf01@telco	Subscriber ID
Download_Kbps	28044	Download rate
Upload_Kbps	10456	Upload rate
Latency	60	Latency rate
Jitter	8	Jitter rate
Packet_Loss	-1	% packet loss
Package	30m	Subscriber pkg
Network_Link	#im01.ppg_kbu#telco#ae0.102#inm_g003 pon 1/1/02/01:10	Node information

The collected test results are compressed and consolidated into a central repository. Regularly, the Compressed File Archive (Bz2) format [1] is used to reduce the storage space [1] and also to maintain the compatibility with other file systems. Some naming conventions were applied for each result set. The average size of the overall test result files is as small as 2.0 KB. The central repositories contain a huge number of small files generated every second. This provision explains the big data characteristic for this telco data set is 'Velocity' and 'Volume'.

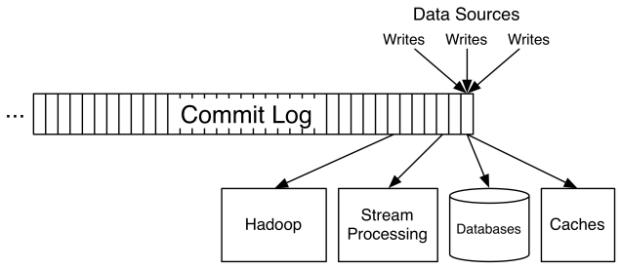

Figure 12. Apache Kafka Integration.

Hadoop has issues with small files [3]. Overwhelming with a lot of small files will increase the Hadoop Metastore utilization and the NameNode memory heap size. Thus, this restriction is not efficient for Hadoop to work. Hadoop Metastore is the central repository that stores all table structures and the NameNode maintains the location reference of the files. The recommended size to store a single file inside Hadoop must be higher than default block size which is 64 MB (Hadoop 1.0) and 128 MB (Hadoop 2.0). Apache Kafka comes into the picture to solve the small files issues by extracting the Bz2 files into the data pipelines. Kafka Consumer will automatically consolidate the small file inside HDFS.

4.4 Subscriber Network Data

An example of network data set is from the Switching and Forwarding Module (SFM). It is capable processing of near 40 million event packets per seconds. The objective of this operation is to capture inbound and outbound network packet of network alarm. This activity generates huge amount of data at high velocity. The related characteristics of this data set are 'Velocity' and 'Volume'.

Table 7. SFM Alarm Data set Column Example

Column Name	Value	Description
Messageid	Vano-B-OI_Alu Ftth Event_ JUP_G005	Unique Message Identification
Notificationkey	JUP_G005_1482072	Notification Key
Messagetype	Realtime	Message Type
Domain	Vano-B-OI	Domain Name
Name	JUP_G005_	MSAN Name
Class	Alu Ftth Event	MSAN Class
Event	Dg	Event CodeName
Elementclass	Olt	Element Class
Elementname	JUP_G005	Element Name
Severity	3	Severity Level
…	…	

The raw network data set structure mostly is in a semi-structured format. To make this format readable, it requires a transformation using a programming language or tools. R language is an example language used to extract the message from the semi-structured format such as JavaScript Object Notation (JSON) or Extensible Markup Language (XML). The post processing messages are consolidated through the Apache Kafka and written back to HDFS.

4.5 Subscriber Social Media Data

The interaction with the social media had dramatically changed in the past few years. Previously, it was used for sharing information online and online discussion. Today, telcos utilize the social media as one of the communication channels with their active subscriber and as well as for the potential customers. Now, online social networks such as Twitter and Facebook are considered by the community as the medium to write personal opinion about any issues regarding of telco services and products. Due to ease of accessibility, many telcos have utilized the social media to engage their customer officially.

Managing the social media channel has become one of the challenging jobs especially for the telco's marketing department. New personal messages and complaints from the subscriber are critical to the company. Telco uses the social media API to extract the complaint messages in real-time for faster response to the problem. Three characteristics classified for telco social media data are 'Velocity,' 'Volume,' and 'Variety.'

```
library("twitteR")
library(methods)
   setup_twitter_oauth("kcCCiP1cdcewQad3CNiwv5QjDYoZf",
    "P6WtWIzgrgPeeeU3bKzB3zlOKhL50Q6EJSFt49zlEUyKXqBkBnp01",
   access_token="723ew5111897250406cew41-ZESHiNiPOHElw36Dk1qiLTkRycWFEGB",
   access_secret="ybmO63DhVt7ciTceeVHH9YaI78wMIx3qnIWLdQg6c3RdgLyc")
```

Figure 13. R integration with Twitter API

The online social media data set needs to downstream for further proactive action. The R package called 'twitteR' implements Open Authorization (OAuth) for token based authorization [38] between R and the Twitter server. Apache Kafka is utilized for pipelining Twitter messages and consolidated inside Hadoop HDFS.

4.6 Subscriber Wi-Fi Data

Public Wi-Fi services offer the best digital experience for their subscriber and the potential customer. By collaborating with the other businesses such as fast food shop, franchise, and restaurant, it can boost their sales and earn more customers loyalty. This collaboration improves the marketing strategy by allowing custom logo and new product information on the Wi-Fi landing page.

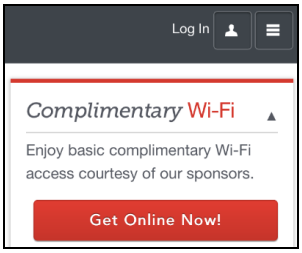

Figure 14. Wi-Fi landing page.

For the food franchise owner, the customer may stay longer and can help to promote the places with their family and friends. The raw Wi-Fi data set is collected from central repositories from multiple locations of the Wi-Fi access point. Each access point managed by walled-garden software to prevent unauthorized access. The average size of each raw files is around 15 GB daily per access point. The files contain all the customer's activities, for example, how long they spend online, and their total downloads or uploads. The characteristic of Wi-Fi data set can be classified as 'Velocity' and 'Volume.' The following table shows the sample columns for the Wi-Fi data set.

Table 8. WIFI Data Example

Column Name	Value	Description
Username	6c8dc1266dc8	User login
Starttime	2016-08-18 08:03:15	Session start
Endtime	2016-08-18 08:08:30	Session end
Duration	15	In seconds
Download	0.0016584843	In Gigabytes
Upload	0.0019236129	In Gigabytes
Apname	nam_001_pd0045_pa001	access point
Apgroup	telcowifi-apg	Group access
Userrole	wallgarden	Role
SSID	telco wifi	SSID Name

Due to multiple locations of the access point, a specific tool is required to support the storage dimension and size. It needs to take a few preprocessing steps before it can access Business Intelligence (BI) tools for generating the reports. The data ingestion process can be utilized inside Hadoop via the data management components such as StreamSets. StreamSets is an open source enterprise-grade of big data ingestion infrastructure with interactive command and control. Kafka data-flow ingestion operates inside the StreamSets component toward continues data movement [39] and integration between the Wi-Fi data and HDFS.

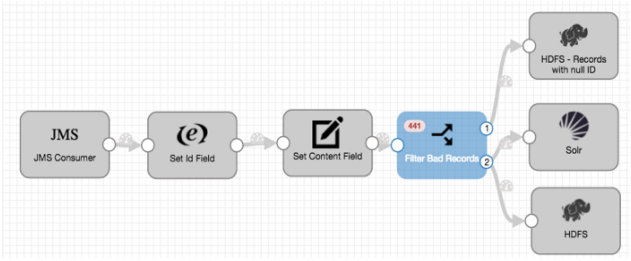

Figure 15. StreamSets Workflow Example

The next stage is aggregating of the Wi-Fi data set using Apache Hive. Hive works as data refinery and aggregator [40] for Hadoop ecosystem. Apache Spark can work on the top of Hive as an alternative to MapReduce framework for faster processing. The following example SQL script work with Spark framework for aggregating the total number of access point of *telco_translog_wifi* table inside the HDFS.

```
set hive.execution.engine = spark ;
select access_point , count(*) as total_ap from
telco_translog_wifi group by access_point ;
```

Apache Spark works as the execution engine and utilizes computing resources such as YARN. This approach will improve the performance and query planning for Hadoop. The next step is, converting the raw Wi-Fi data set into a Parquet columnar format. This columnar format can improve query performance by the external BI tools via interfaces such as JDBC and ODBC.

5. Conclusion

The convergence of big data in telco industry is shaping the future of telco how to drive business value from data analytics capabilities. Digital era has moved data accessibility from batch into real-time. The capability to access large volume inside big data platform helps the data scientist to produce meaningful analytics results. Big data analytics platform has enabled the data scientist to work with the massive amount of data set without restriction. This is why many telcos have moved from hypothesis based towards data driven approach. Big data platform enables the environment which encourages data discovery. As a result, telco now can move faster, do more experiments and learn quickly. Now, they can load all the data and let the data tell the story.

References

[1] E. R. Schendel, A. M. Mahdy, "Archiving with Athamas: A Framework for Optimized Handling of Domain Knowledge", *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*, doi:10.1109/dbkda.2010.30Ssq, 2010.

[2] F. Wang, "A High Performance Architecture for Web Service Systems in XML", *2007 International Conference on Service Systems and Service Management*, doi:10.1109/icsssm.2007.4280189, 2007.

[3] C. Vorapongkitipun, N. Nupairoj, "Improving performance of small-file accessing in Hadoop", *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, doi:10.1109/jcsse.2014.6841867, 2014.

[4] A. Sheth, "Transforming Big Data into Smart Data: Deriving value via harnessing Volume, Variety, and Velocity using semantic techniques and technologies", *2014 IEEE 30th International Conference on Data Engineering*, doi:10.1109/icde.2014.6816634, 2014.

[5] R. I. Jony, A. Habib, N. Mohammed, R. I. Rony, "Big Data Use Case Domains for Telecom Operators", *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, doi:10.1109/smartcity.2015.174, 2015.

[6] A. Lheureux, K. Grolinger, H. F. Elyamany, M. A. Capretz, "Machine Learning With Big Data: Challenges and Approaches", *IEEE Access*, vol. 5, pp. 7776-7797, doi:10.1109/access.2017.2696365, 2017.

[7] S. Dwivedi, P. Kasliwal, S. Soni, "Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime) ", *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, doi:10.1109/cdan.2016.7570894, 2016.

[8] A. N. Richter, T. M. Khoshgoftaar, S. Landset, T. Hasanin, "A Multi-dimensional Comparison of Toolkits for Machine Learning with Big Data", *2015 IEEE International Conference on Information Reuse and Integration*, doi:10.1109/iri.2015.12, 2015.

[9] C. Stergiou, K. E. Psannis, "Algorithms for Big Data in Advanced Communication Systems and Cloud Computing", *2017 IEEE 19th Conference on Business Informatics (CBI)*, doi:10.1109/cbi.2017.28, 2017.

[10] A. Pal, S. Agrawal, "An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using HDFS and MapReduce", *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, doi:10.1109/cnsc.2014.6906718, 2014.

[11] S. Zhao, M. Chandrashekar, Y. Lee, D. Medhi, "Real-time network anomaly detection system using machine learning", *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, doi:10.1109/drcn.2015.7149025, 2015.

[12] "Spark Streaming Programming Guide", (n.d.). Retrieved August 28, 2017, from https://spark.apache.org/docs/latest/streaming-programming-guide.html

[13] F. Zhu, J. Liu, S. Wang, J. Xu, L. Xu, J. Ren, …, T. Huang, "Hug the Elephant: Migrating a Legacy Data Analytics Application to Hadoop Ecosystem", *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, doi:10.1109/icsme.2016.14, 2016.

[14] Y. Xiaoshan, Z. Ligu, Z. Qicong, F. Dongyu, "Research on Evaluation Method of Big Data Storage Utilization", *2016 4th Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science &*

Engineering (ACIT-CSII-BCD), doi:10.1109/acit-csii-bcd.2016.077, 2016.

[15] "Google Hangouts vs. Skype: A comparative look", (n.d.). Retrieved August 28, 2017, from http://searchunifiedcommunications.techtarget.com/feature/Google-Hangouts-vs-Skype-A-comparative-look

[16] "Top 3 Data Challenges for Telcos: Escaping Silos to Pursue Scale", (2017, July 25). Retrieved August 28, 2017, from https://www.thinkbiganalytics.com/2016/12/07/escaping-silos-pursue-scale-top-three-big-data-challenges-telcos/

[17] (n.d.). Retrieved from http://www.actuate.com/download/casestudy/Telecom-Hadoop-Case-Study.pdf

[18] (n.d.). Retrieved from http://www.huawei.com/ilink/en/download/HW_323807

[19] (n.d.). Retrieved from http://bigdata-madesimple.com/11-interesting-big-data-case-studies-in-telecom/

[20] (n.d.). Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-da-ba-making-the-right-call-062613.pdf

[21] (n.d.). Retrieved from https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/Telecoms/PDFs/RecallNo21_Big_Data_2012-07.ashx

[22] (n.d.). Retrieved from https://conferences.oreilly.com/strata/big-data-conference-sg-2015/public/schedule/detail/45145

[23] (n.d.). Retrieved from http://www.ey.com/Publication/vwLUAssets/EY_-_Optimize_network_OPEX_and_CAPEX_while_enhancing_the_quality_of_service/$FILE/EY-optimize-network-opex-and-capex.pdf

[24] (n.d.). Retrieved from https://www.linkedin.com/pulse/big-data-telecommunications-vijay-gaur

[25] (n.d.). Retrieved from https://www.singtel.com/content/dam/singtel/business/globalservices/Featured%20Articles/Singtel%20Endpoint%20Threat%20Detection%20and%20Response.pdf

[26] (n.d.). Retrieved from https://www.ericsson.com/res/docs/2012/capitalizing_on_customer_experience.pdf

[27] "New Service Development Process in Telecom Industry: The ... " (n.d.). Retrieved August 28, 2017, from http://www.diva-portal.org/smash/get/diva2:678791/FULLTEXT01.pdf

[28] N. Miloslavskaya, A. Tolstoy, "Application of Big Data, Fast Data, and Data Lake Concepts to Information Security Issues", *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, doi:10.1109/w-ficloud.2016.41, 2016.

[29] S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, …, P. Poulosky, "Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming", *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, doi:10.1109/ipdpsw.2016.138, 2016.

[30] G. Valkanas, T. Lappas, D. Gunopulos, "Mining Competitors from Large Unstructured Datasets", *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1971-1984, 2017.

[31] (n.d.). Retrieved August 28, 2017, from https://mahout.apache.org/users/basics/algorithms.html

[32] "KNIME Analytics Platform", (n.d.). Retrieved August 28, 2017, from https://www.knime.com/knime-analytics-platform

[33] "What is R? " (n.d.). Retrieved August 28, 2017, from https://www.r-project.org/about.html

[34] R. (n.d.). RevolutionAnalytics/RHadoop. Retrieved August 28, 2017, from https://github.com/RevolutionAnalytics/RHadoop/wiki

[35] S. Kumar, P. Singh, S. Rani, "Sentimental analysis of social media using R language and Hadoop: Rhadoop", *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, doi:10.1109/icrito.2016.7784953, 2016.

[36] R. Devakunchari, "Handling big data with Hadoop toolkit", *International Conference on Information Communication and Embedded Systems (ICICES2014)*, doi:10.1109/icices.2014.7033839, 2014.

[37] A. Medem, M. Akodjenou, R. Teixeira, "TroubleMiner: Mining network trouble tickets", *2009 IFIP/IEEE International Symposium on Integrated Network Management-Workshops*, doi:10.1109/inmw.2009.5195946, 2009.

[38] (n.d.). Retrieved from https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

[39] "What is Dataflow Performance Management? " (n.d.). Retrieved August 28, 2017, from https://streamsets.com/data-performance-management/

[40] A. (2017, August 28). Apache/hive. Retrieved August 28, 2017, from https://github.com/apache/hive