ISSN No:-2456-2165

Telco Big Data Analytics using Open-Source Data Pipeline: Use Cases, Detailed Use Case Implementation Results and Findings

Dr. Chandrasekar Shastry and Abirami Thangavel

Abstractl:- Operators of telecommunications are sitting on a gold mine. They produce enormous amounts of data each day, up to billions of CDRs and events. These data could be user, network, or customer-related. For the telecommunications operators, effectively gathering, storing, processing, and analyzing this amount of data can be very difficult. The infrastructure must have ample storage space and computational power. Additionally, it needs adaptability to assess various data formats. Therefore, it is crucial to create the best architecture possible in order to overcome these technical difficulties and satisfy commercial needs.

In this paper, we have used the seven layers of implementation described in the previous work and implemented a potential use case- churn analysis of telecom customers. We have also analyzed various other use cases along with case studies and have proved how our open source data pipeline architecture would help the telecommunication sectors to implement and analyze those use cases.

Keywords:- *Implementation of Open Source Data Pipeline for BDA, Customer Churn Analysis, Churn Analysis Code, Big Data Analytics, Telecom Use Cases for BDA, Open Source, Lambda Architecture, Big Data Architecture Layers, Telecommunication, Telecom BDA Use Cases.*

I. INTRODUCTION

The telecommunications sector collects a huge amounts of data for various decision-making and business needs. A surge in the amount of data flowing across telecom operatornetworks has been brought on by the quick increase in the use of smartphones and other connected mobile devices. The operators must organize, process, and gather knowledge fromthe data that is already available. By assisting in theoptimization of network utilization and services, improving customer experience, and enhancing security, big data analytics can help them maximize profitability. According to research, there is a significant chance that big data analytics will help telecom firms.

Big Data's potential presents a dilemma, though: how can a business use data to boost sales and profitability across the whole value chain, including network operations, product development, marketing, sales, and customer service. For instance, Big Data analytics helps businesses to forecast peak network demand so they can take action to reduce congestion. Additionally, it can assist in identifying consumers who are most likely to experience financial difficulties paying their bills as well as those who are set to switch operators, hence escalating churn.

When it comes to Big Data analytics, operators are typically warned against using the traditional top-down strategy, which identifies the problem that needs to be solved before looking for the data that might be able to help. The operators should instead concentrate on the data itself, using it to draw linkages and correlations. If used properly, the data could produce insights that could serve as the foundation for more efficient operations.

Before beginning any BDA venture, it's crucial to determine the obstacles that can prevent project execution as well as the advantages that will be available once the solution isimplemented. We discussed the main advantages and disadvantages of big data deployment in the telecom sector in this part.

The deluge of data produced by linked devices, consumer behavior, social media networks, call data records, government portals, and billing information is posing challenges for telecom providers. Based on their research into the use of big data analytics in South Africa, I. Malaka and I. Brown divided the problems into three categories: technological, organizational, and environmental (Malaka and Brown, 2015). Only a few significant use case implementations—which, in my opinion, have the greatest bearing on the BDA's implementation—will be covered in this evaluation.

We conducted a rigorous procedure of assessing the literature in order to study the aforementioned research concerns. The project and data governance approaches, frameworks, and skill requirements of the BDA are specifically addressed in this assessment. Then, in our earlier works, we detailed and examined the most popular methodologies and architectural designs already in use by a number of telecom carriers, as well as the relevant talents required to make such projects effective. In the last half of this paper, we also make suggestions for additional prospective use cases. In this article, we provide implementation examples of various sample use cases that have been successfully implemented.

II. TELECOM USE CASES FOR CASE STUDY

A. Customer Churn Prevention

One of the most well-liked BDA use cases created in the telecom industry is the churn prediction. This is because acquiring a new customer is more expensive than keeping a current one, which is why. The topic has been covered in a number of research publications from various perspectives. The Hidden Markov Model was used to construct the first use case that we quote (Xia et al., 2018). The training set

ISSN No:-2456-2165

was made up of five months' worth of historical customer, billing, and network data. The application was created using a telecom operator's Hadoop Platform. For operations on statistical data, Hive was employed. The churn application was created by the project team utilizing these two other well- known classifiers for the purpose of comparison: Rough Forest and LIBLINEAR.

The second example of preventing churn is a little more novel since it applies modelling techniques from the health industry to the telecommunications industry (Kurt et al., 2019). The authors wanted to underline how crucial transdisciplinary abilities are for data science. The authors discovered parallels between survival and churn. For forecasting telecom churners, the survival vector machine method was chosen. With an AUC of 0.82, the final application did fairly well.

The case of SyriaTel (Ahmad et al., 2019), in which the authors employed the Extreme Gradient Boosting "XGBOOST" model to develop the churn prevention application, is another intriguing churn scenario to mention. The team employed customer social network data in addition to the traditional data often used for churn prediction (CRM, Billing, and Network KPI) to attain a better performance. Three other algorithms were tested by the authors: Gradient Boosted Machine Trees ("GBM"), Random Forests, and Decision Trees. However, using the XGBOOST algorithm, which attained an AUC of 0.933, produced the best results.

As a last illustration, consider a system that Pakistani researchers (Khan et al., 2019) created employing Deep Learning techniques, particularly Artificial Neural Networks Multiple attributes, including demographic information from CRM.

B. Offer Tendency

It is possible to provide tailored offers or services that meet customers' needs by combining analytics techniques with usage traffic, loyalty points, event-based promotion, and demographic data. In order to analyze client preferences and spot business possibilities, a telecom operator in the Asia- Pacific area (Fox, 2015) launched a predictive analytics solution that generates propensity models. The operator achieved exceptional outcomes by increasing efficiency and competitiveness, increasing the speed of its ad hoc reporting by 190x, and realizing 10% higher net revenues.

C. Preventing Revenue Leakage

Telecom operators have seen a variety of revenue leakages during the last few years. One of the most common was SIM box fraud. Government and the telecoms industries suffered severe losses. It was projected that throughout Africa, it averaged 150 million US dollars yearly. A technique used in telecommunications called SIM boxing involves setting up hardware that can accommodate many SIM cards and using it to end international conversations that have been made using voice over IP. Two methods for SIM Box identification were mentioned by Chung-Min Chen: proactive test calls and passive CDR analysis (Chen, 2016). The first technique needs the operator to place calls to its own country from other countries and determine the kind of termination.

Because it needs the operator to provide complete coverage in order to catch all potential fraudsters, this strategy is too labor-intensive to implement. However, traffic analysis will be far less expensive and guarantee better outcomes. Because of this, a few rules have been established to identify suspect SIM cards based on the volume and location of calls made.

D. Improvement of Customer Experience

Big Data capabilities and analytical tools have completely changed how telecom carriers handle their consumer relationships. As an illustration, some operators created services to notify their clients if a network issue arises. The number of calls that the call centers got decreased as a result. Using chatbots, further projects aimed at enhancing the customer experience were created. Results were observed at several levels, including Opex optimization, an increase in Net Promoter Score (NPS), and higher employee satisfaction as a result of their redistribution to more motivating tasks. A study about the application of real-time customer experience prediction for a large telecommunications operator in Africa was conducted by E. Diaz-Avile et al (Diaz-Aviles et al., 2015). Without directly interacting with the end user, the authors acknowledged that it is difficult to predict the sort of user experience (good or unpleasant) at any given time. As a result, the writers' guiding premise for the system was that any negative customer experiences would lead to calls to the telco's care center.

E. Taking Proactive Measures

Big data solutions have been developed by a number of telecommunications companies to proactively identify network issues before they have an impact on end users. These solutions make it possible, among other things, to offer the Telecom operator the proper suggestions, enabling them to take action in advance and prevent any effects on sales and customer satisfaction.

F. Mobile Location Information to Prevent COVID 19

Telecom operators have taken various measures during the pandemic to effectively spread awareness and to help in preventing the spread of the disease. Mobile Location Data for COVID-19 Prevention During the COVID-19 pandemic period, governments of several countries (Doffman, 2020) in Asia, Europe and the United States of America have developed solutions based on mobile users anonymized location data, to track their movements, in order to control and limit the spread of the Corona virus.

These solutions enabled them, on one hand, to track people infected by the Covid-19 by retracing their travel routes, places visited, as well as people met, and on the other hand, highlighted the areas where people didn’t respect public health safety measures.

ISSN No:-2456-2165

G. Multi-SIM Detection with Social Network Analysis

In order to continuously increase their earnings, certain telecom operators have developed analytical tools based on the social network theory that enable them to visualise the connections between the various network users. This aided them in locating the most influential clients so they could target them with special offers. In the N. R. Al-Molhem et al. example study, using SNA to target influencers led to a 30% increase in mobile traffic above conventional methods (Al- Molhem et al., 2019). The authors of the study have created a tool to track customers switching between numerous SIM cards. The concept of "mutual friends" served as the foundation for the approach.

To find the SIM pairs that might be the same user's representation, two scores have been computed. The original one used a similarity score based on the Cosine and Jaccard measurements. The second was a behavioural score derived from an examination of CDRs. The application's accuracy rate for clients of the same operator was 92%.

H. Public Security BDA

Some governments utilise BDA in conjunction with cloud services and IoT technology to increase the security of its residents. This is accomplished by forecasting the locations where crimes are most likely to occur at any given time. By combining public security data with telecoms data, this has been made possible. The "intelligent police" watched as insights were provided to enable suspect monitoring, crime early detection, crime kinds, case analysis, and clue extraction. When it comes to the tools, W. LIANG et al. initially used the Xgboost method to simulate the answer (Liang et al., 2018). The accuracy gained was just about 48%, hence the outcome was unsatisfactory. The authors then made the decision to investigate additional perspectives on realisation and used the Self Excitation Point Process Model.

I. Power Saving

Energy conservation in data centres (DC) is a significant topic that data scientists have been working on lately. The finest case study is that of J. Gao, who used AI techniques to optimise energy utilisation in Google Data Centers (GAO, 2014). W. LIANG et al. used deep learning (DL) techniques to address the problem of data centre power reduction in China, and their approach was successful (Liang et al., 2018). For the purpose of forecasting the trend in energy use, a five layer DL model was used. About 19 elements made up the model variables, such as the overall server IT load, resource utilisation, the environmental index of the DC (temperature and humidity), etc. The following phases form the foundation of the reasoning behind the creation of the energy-saving application:

- Keeping track of the workloads of all virtual machines (VMs) operating on all real hardware.
- Finding the virtual machines with the fewest jobs and the lowest load.

J. Real-time Traffic Analysis

Telecom operators must guarantee extensive radio coverage of all the locations where their customers and potential customers are present in order to provide impeccable service quality. The IP backbone infrastructure and the network equipment used for this purpose are producing a significant amount of data that is essential for effective real-time traffic control. In contrast to the crowd sourced technique, the authors offered a novel strategy based on micro-level traffic modelling that also took into account the traffic's temporal dimension. This made it possible to monitor traffic at various times. Among the fundamental considerations for the solution design were customer data protection and the cost effectiveness of the solution.

III. SAMPLE USE CASE IMPLEMENTATION: CHURN ANALYSIS

The dataset is taken from the following:https://www.kaggle.com/becksddf/churn-in-telecoms- dataset/data. We have considered a huge volume of customerdata for wider analysis and prediction. This data covers wide range of area code, international customers, local customers,and of various plans.

The following code snippet displays the list of all thelibraries imported.

```
from sklearn import cross_validationfrom sklearn import tree
from sklearn  import  svm from sklearn import ensemble
from sklearn import neighbors from sklearn import linear_modelfrom sklearn import metrics
from sklearn import preprocessing

%matplotlib inline
from IPython.display import Imageimport matplotlib as mlp
import matplotlib.pyplot as pltimport numpy as np
import os
import pandas as pd import  sklearn import seaborn as sns


#df.dtypes
```

ISSN No:-2456-2165

The following code snippet reads the input from the .csv file. The input is the dataset that is used for further processingand churn analysis and prediction.

```
#df = pd.read_csv('../input/mytest.csv')
df =
pd.read_csv('../input/bigml_59c28831336c6604c
800002a.csv')
print (df.shape)
```

The following code snippet loads the data. The results aredisplayed in the Results section.

First conclusion is that the data are unbalanced since thereare fewer data points in the True Churn group. The following method describe () calls the methodimplementation.

```
df.describe()
```

Various types of analysis are captured as follows:

A. State-wise Customer Churn

The following code snippet shows how to filter the churnedcustomers state-wise. The results are captured in the next section.

```
df.groupby(["state",
"churn"]).size().unstack().plot(kind='bar',
stacked=True, figsize=(40,10))
```

A. Area-wise Customer Churn

The following code snippet shows how to filter the churnedcustomers area-wise. The results are captured in the next section.

```
df.groupby(["area code",
"churn"]).size().unstack().plot(kind='bar',
stacked=True, figsize=(5,5))
```

```
y = df["churn"].value_counts()
#print (y)
sns.barplot(y.index, y.values)
y_True = df["churn"][df["churn"] == True]
print ("Churn Percentage = "+str(
(y_True.shape[0] / df["churn"].shape[0]) *
100 ))
```

```
# Load data
df.head(3)
```

ISSN No:-2456-2165

B. Plan-wise Customer Churn (International Plan)

The following code snippet shows how to filter the churned customers with an international plan. The results are captured in the next section.

```
df.groupby(["international plan",
"churn"]).size().unstack().plot(kind='bar',
stacked=True, figsize=(5,5))
```

C. Plan-wise Customer Churn (Voice Mail Plan)

The following code snippet shows how to filter the churnedcustomers with voice mail plan. The results are captured in the next section.

```
df.groupby(["voice mail plan", "churn"]).size().unstack().plot(kind='bar', stacked=True, figsize=(5,5))
```

The following code snippet shows how to managecategorical columns and label encoding:

```
# Discreet value integer encoderlabel_encoder =
preprocessing.LabelEncoder()
# State is string and we want discreetinteger values
df['state'] = label_encoder.fit_transform(df['state'])
df['international plan'] = label_encoder.fit_transform(df['internationalplan'])
df['voice mail plan'] = label_encoder.fit_transform(df['voice mailplan'])

#print (df['Voice mail plan'][:4])print (df.dtypes)
df.shape df.head()
```

The following code snippet is used to strip response valuesand redundant columns:

```
#Strip Response Values
y = df['churn'].as_matrix().astype(np.int)y.size

#Strip Redundant Columns
# df = df.drop(["Id","Churn"], axis = 1,inplace=True)
df.drop(["phone number","churn"], axis =1, inplace=True)
```

The following code snippet provides cross validation:

```
def stratified_cv(X, y, clf_class, shuffle=True, n_folds=10, **kwargs):
    stratified_k_fold = cross_validation.StratifiedKFold(y,n_folds=n_folds, shuffle=shuffle)
    y_pred = y.copy()# ii -> train
    # jj -> test indices
    for ii, jj in stratified_k_fold: X_train, X_test = X[ii], X[jj]y_train = y[ii]
```

ISSN No:-2456-2165

IV. RESULTS

A. *Churn Over all Ratio*

The Total number of custom ers who have churned is represented in as “True” (Orange color). Boolean variable is used to get an analysis of total ratio of churned customers to the customers who have retained (represented in blue).

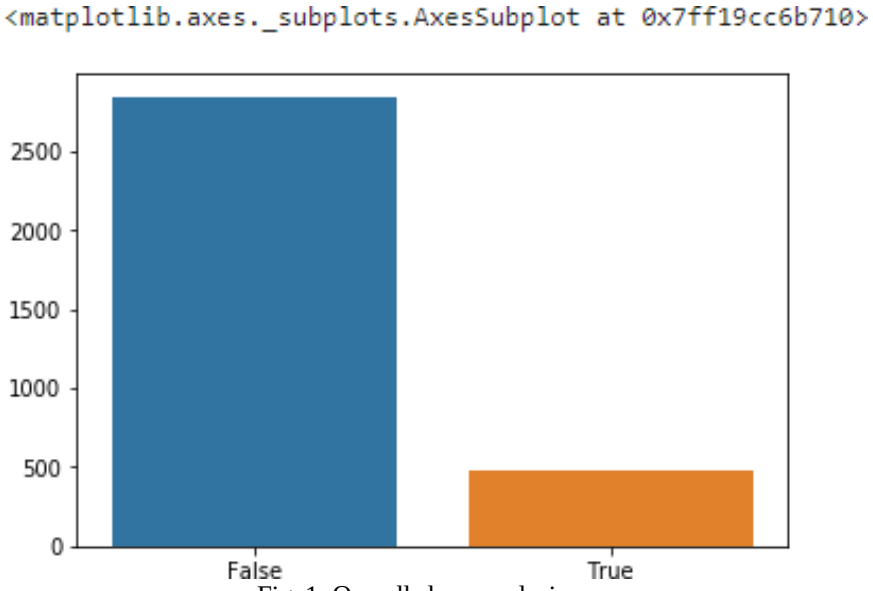

Fig. 1: Overall churn analysis

B. *Customers Churn Based on Area Code*

The customer churn is calculated with respect to the area code. In the following chart, the number of customer churn is represented in orange.

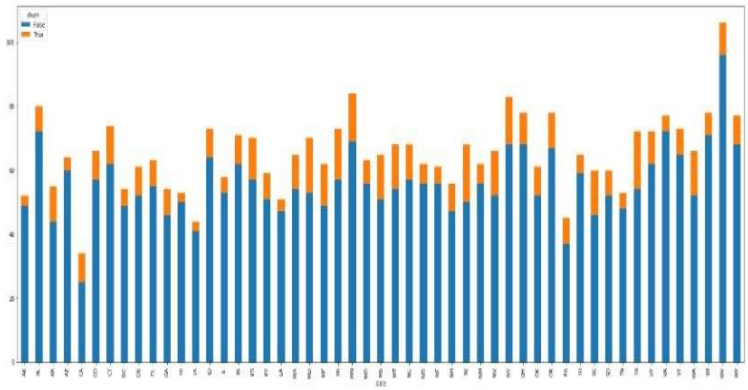

Fig. 2: Overall churn analysis

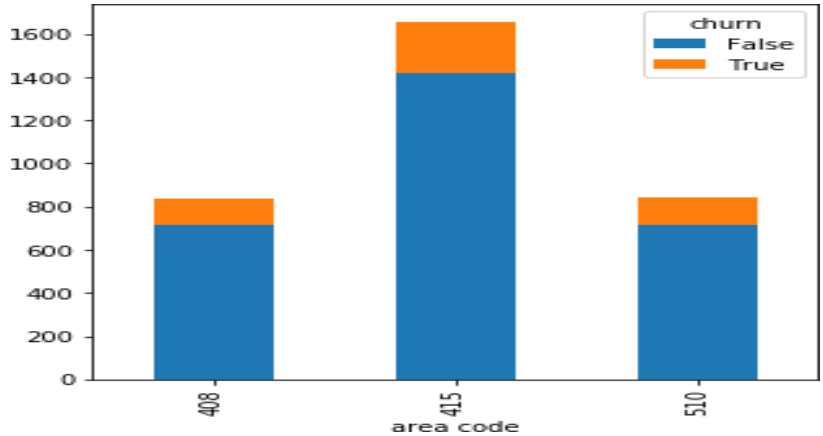

Fig. 3: Customers churn based on area code

ISSN No:-2456-2165

C. Customers Churn Based on International Plan

The total number of churned customers who were using aninternational plan is represented in orange.

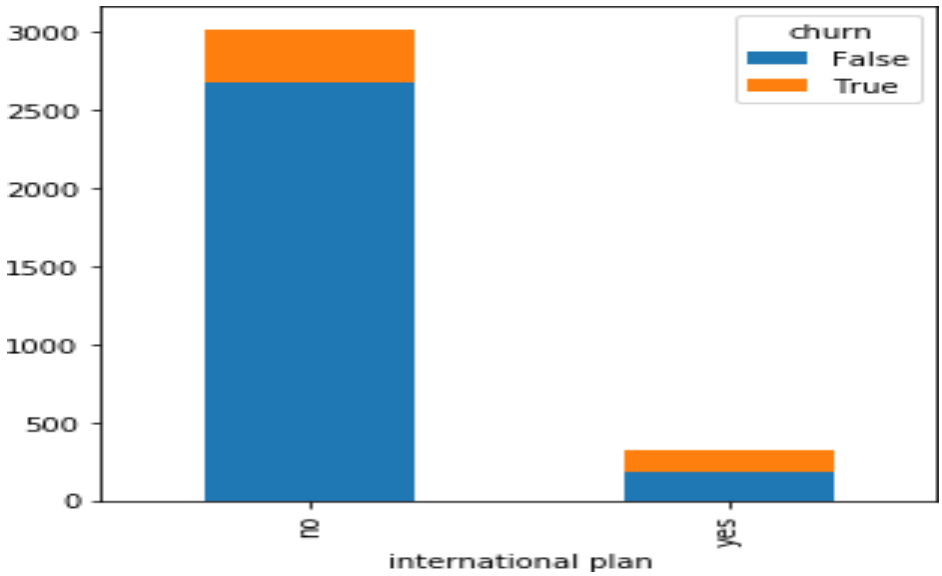

Fig. 4: Customers churn based on plan-international plan

D. Customers Churn- Voice Mail Plan

The total number of churned customers who were using avoice mail plan is represented in orange.

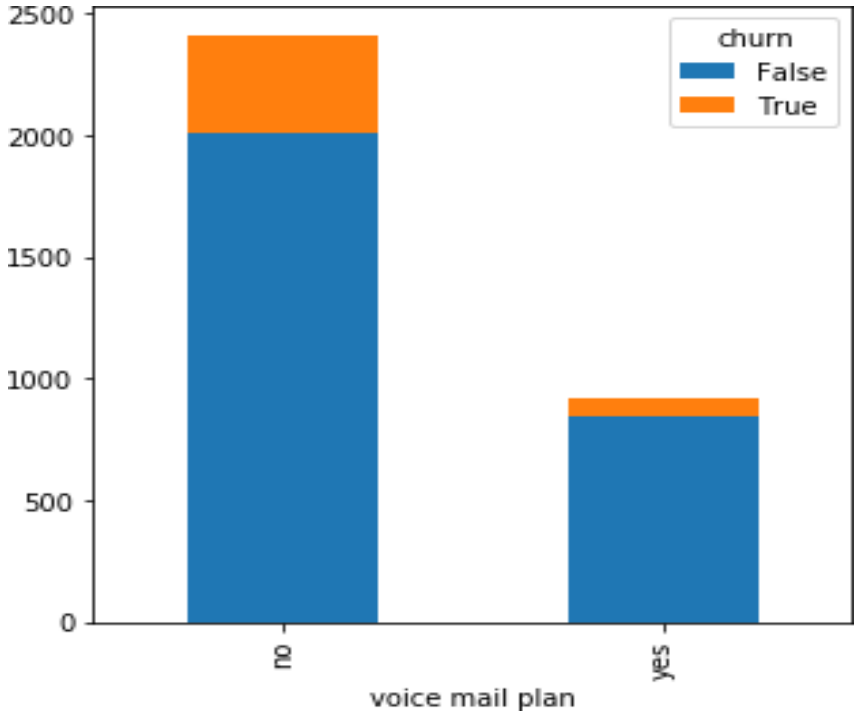

Fig. 5: Customers churn based on plan-voice mail plan.

E. Additional Data and Deatiled Churn Results

Customer churn based on various categories such as area code, and plan are calculated individually and displayed in the form of charts as displayed above. The following tables display the overall results in a more descriptive way. Along with customer churn, total number of calls in minutes, days, total calls made in the evening, total calls made in the night, state-wise data, customer service calls made along with additional details are captured and reported.

ISSN No:-2456-2165

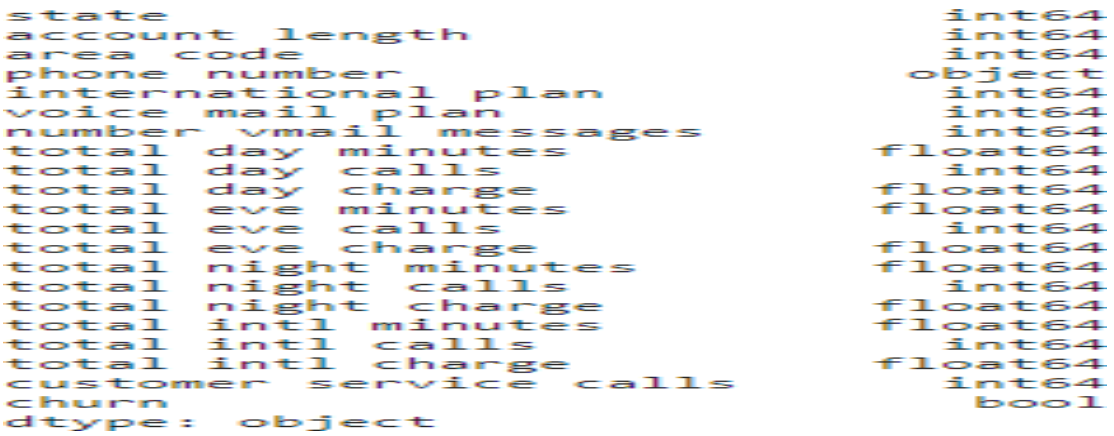

Fig. 6: Overall churn analysis

area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
415	382-4657	0	1	25	265.1	110	45.07	...	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
415	371-7191	0	1	26	161.6	123	27.47	...	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
415	358-1921	0	0	0	243.4	114	41.38	...	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
408	375-9999	1	0	0	299.4	71	50.90	...	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
415	330-6626	1	0	0	166.7	113	28.34	...	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False

Fig. 7: CuInstdoimviedrusaclhcursntobmaseerdeotnaiplslawn-ivthoiacedemtaiilledplcahnu. rn analysis table 1.

area code	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls
3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

Fig. 8: Individual customer details with a detailed churn analysis table 2.

V. CONCLUSION AND FUTURE WORK

BDA innovations have transformed the era of communications. The Hadoop ecosystem, streaming analytics tools, and open source tools gave telecom operators new waysto mine previously untapped data sets for insights. The bestapproaches for controlling the project and the data must firstbe defined in order to profit from BDA solutions. Second, pick an architecture that will take into account all thepeculiarities and demands of the telecom industry. Thisincludes the capacity to offer real-time insights and the processing of batch and streaming data. In this paper, wehave listed a detailed case study of various use cases thatwould benefit from our architecture and implementation. Finally, we have concluded with a detailed implementation ofa sample use case: churn analysis using Python code, opensource tools and have derived the results. We have done a detailed analysis and captured results along with the findings. As a future work, we want to produce proof of concepts based on the Kappa+ architecture and create the identical use cases for both configurations in order to assess and contrast the outcomes.

ISSN No:-2456-2165

REFERENCES

[1.] A.K. Ahmad., A. Jafar, and E. Aljoumaa,. "Customer churn prediction in telecom using machine learning in Big Data Platform," Journal of Big Data, vol. 6, no. 1, 2019. Availablet: https://doi.org/10.1186/s40537-019-0191-6..

[2.] Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, "Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study," *IEEE Network*, vol. 30, no. 2, pp. 54–61, 2016.

[3.] X. Diebold, "Big data dynamic factor models for macroeconomic measurement and forecasting," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, "(edited by M. Dewatripont, LP Hansen and S. Turnovsky)*, 2003, pp. 115–122.

[4.] B. Violino, "How to avoid big data analytics failures," https://www. infoworld.com/article/3212945/big-data/how-to-avoid-big-data- analytics-failures.html, 2017.

[5.] P. Zikopoulos, C. Eaton *et al*., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 2011.

[6.] Demirkan and B. Dal, "The data economy: Why do so many analytics projects fail?" [Online]. Available: http://analytics-magazine.org/the-data-economy-why-do-so-many- analytics-projects- fail/, 2014.

[7.] M. E, "The world according to linq," *Communications of the ACM*, vol. 10, no. 54, pp. 45–51, 2011.

[8.] J. Ohlhorst, *Big Data Analytics: Turning Big Data into Big Money*, 1st ed., Amazon, Ed. John Wiley & Sons, 2012.

[9.] C. M. Ricardo and S. D. Urban, *Databases Illuminated*, 3rd ed., Amazon, Ed. Jones & Bartlett Learning, 2015.

[10.] C. Deka, *NoSQL: Database for Storage and Retrieval of Data in Cloud*, Amazon, Ed. Chapmanand Hall/CRC, 2017.

[11.] M. D. D. Silva and H. L. Tavares, *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*, 2nd ed., Amazon, Ed. O'Reilly Media, 2013.W.-K.

[12.] Datafloq, "Top reasons of hadoop - big data project failures," https:// datafloq.com/read/top-reasons-of-hadoop-big-data-project-failures/ 2185, 2017.

[13.] Kumari, S. Tanwar, S. Tyagi, N. Kumar, M. Maasberg, and K.- K. R. Choo, "Multimedia big data computing and internet of things applications: A taxonomy and process model," J. Network and Computer Applications, vol. 124, pp. 169–195, Dec. 2018.

[14.] Manyika, M. Chui, M. G. Institute, B. Brown, J. Bughin, R. Dobbs, Roxburgh, and A. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey, 2011. [Online].

[15.] Available: https://books.google.com.pk/books?id=APsUMQAACAAJ

[16.] Daki, A. El Hannani, A. Aqqal, A. Haidine, A. Dahbi, and H. Ouahmane, "Towards adopting big data technologies by mobile networks operators: A moroccan case study," in *Proc. 2nd IEEE Int. Conf. Cloud Computing Technologies and Applications*, 2016, pp. 154–161.

[17.] T. White, *Hadoop: The Definitive Guide*, 3rd ed., Amazon, Ed. USA: Yahoo Press, 2012.

[18.] C. M. Murphy, "Writing an effective review article," *Journal of Medical Toxicology*, vol. 8, no. 2, pp. 89–90, Jun 2012. [Online].

[19.] Available: https://doi.org/10.1007/s13181-012-0234-2

[20.] Chih-Lin, Y. Liu, S. Han, S. Wang, and G. Liu, "On big data analytics for greener and softer ran," *IEEE Access*, vol. 3, pp. 3068–3075, 2015.

[21.] Park, H. Gebre-Amlak, B. Choi, S. Song, and D. Wolfinbarger, "Understanding university campus network reliability characteristics using a big data analytics tool," in *Proc. 11th Int. Conf. Design of Reliable Communication Networks*, March 2015, pp. 107–110.

[22.] S. Parise, "Big data: a revolution that will transform how we live, work, and think, by viktor mayer-schonberger and kenneth cukier," *J. Information Technology Case and Application Research*, vol. 18, no. 3, pp. 186–190, Sept. 2016. [Online]. Available: https://doi.org/10.1080/ 15228053.2016.1220197

[23.] D. Sipus, "Big data analytics for communication service providers," in *Proc. 39th IEEE Int. Conv. Information and Communication Technology, Electronics and Microelectronics*, May 2016.

[24.] Bughin, "Reaping the benefits of big data in telecom," *J. Big Data*, vol. 3, no. 1, 2016.

[25.] M. Rathore, A. Paul, A. Ahmad, M. Imran, and M. Guizani, "Highspeed network traffic analysis: Detecting VoIP calls in secure big data streaming," in *Proc. IEEE 41st Conf. Local Computer Networks*, Nov. 2016, pp. 595–598.

[26.] S. Han, C. -L. I, G. Li, S. Wang, and Q. Sun, "Big data enabled mobile network design for 5g and beyond," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 150–157, 2017. [Online]. Available: https://doi.org/10. 1109/mcom.2017.1600911

[27.] F. Hueske and V. Kalavri, *Stream Processing With Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications*, 1st ed., Amazon, Ed. USA: O'Reilly Media, 2018.

[28.] N. Garg, *Learning Apache Kafka, Second Edition*, 2nd ed., Amazon, Ed. USA: Packt Publishing, 2015.

[29.] J.-C. Tseng, H.-C. Tseng, C.-W. Liu, C.-C. Shih, K.-Y. Tseng, C.-Y. Chou, C.-H. Yu, and F.-S. Lu, "A successful application of big data storage techniques implemented to criminal investigation for telecom," in *Proc. 15th IEEE Conf. Asia-Pacific Network Operations and Management Symposium*, 2013, pp. 1–3.

[30.] T. Yigit, M. A. Cakar, and A. S. Yuksel, "The experience of nosql database in telecommunication enterprise," in *Proc. 7th IEEE Int. Conf. Application of Information and Communication Technologies*, 2013, pp. 1–4.

[31.] C. Şenbalcı, S. Altuntaş, Z. Bozkus, and T. Arsan, "Big data platform development with a domain specific language for telecom industries," in *Proc. High Capacity Optical Networks and Emerging/Enabling Technologies*, Dec. 2013, pp. 116–120.

[32.] M. Jonathan and K. Tor, "Subscriber Classification Within Telecom Networks Utilizing Big Data Technologies and Machine Learning," in *Proc. 1st Int. Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pp. 77–84, 2012.

[33.] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Efficient exploration of telco big data with compression and decaying," in *Proc. IEEE 33rd Int. Conf. Data Engineering*, 2017, pp. 1332–1343.

[34.] D. S. Yuri Diogenes, Tom Shinder, *Microsoft Azure Security Infrastructure (IT Best Practices - Microsoft Press)*, 1st ed., Amazon, Ed. USA: Microsoft Press, 2016.

[35.] B. R. Chang, H. F. Tsai, Z.-Y. Lin, and C. -M. Chen, "Access- controlled video/voice over ip in hadoop system with bpnn intelligent adaptation," in *Proc. IEEE Int. Conf. Information Security and Intelligence Control*, 2012, pp. 325–328.

[36.] SolidIT, DB-engines, ranking of key-value stores@ONLINE, 2017. [Online]. Available: https://db-engines.com/en/ranking/key-value+store, https://db-engines.com/en/ranking/document+store, https://db-engines. com/en/ranking/wide+column+store

[37.] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 2785–2792.

[38.] H. Zahid, T. Mahmood, and N. Ikram, "Enhancing dependability in big data analytics enterprise pipelines," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage*, G. Wang, J. Chen, and

[39.] T. Yang, Eds. Cham: Springer International Publishing, 2018, pp. 272–281.

[40.] W. Queiroz, M. A. Capretz, and M. Dantas, "An approach for SDN trafficmonitoring based on big data techniques," *J. Network and Computer Applications*, vol. 131, pp. 28–39, Apr. 2019.

[41.] Zhou, A. Fu, S. Yu, M. Su, and B. Kuang, "Data integrity verification of the outsourced big data in the cloud environment: a survey," *J. Network and Computer Applications*, vol. 122, pp. 1–15, Nov. 2018.

[42.] W. Xu, H. Zhou, N. Cheng, F. Lyu, W. Shi, J. Chen, and X. Shen, "Internet of vehicles in big data era," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 19–35, Jan. 2018.

[43.] Forgeat, "Data processing architectures — lambda and kappa," https:// www.ericsson.com/en/blog/2015/11/data-processing-architectures-- lambda-and-kappa, 2015.

[44.] W. Xu, H. Zhou, N. Cheng, F. Lyu, W. Shi, J. Chen, and X. Shen, "Internet of vehicles in big data era," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 19–35, Jan. 2018.

[45.] ZAGELBAUM, "Kapp. architecture: a different way to process data," https://www.blue-granite.com/blog/a-different-way-to-process-data-kappa-architecture, Jan. 25, 2019.

[46.] S. Jain, M. Khandelwal, A. Katkar, and J. Nygate, "Applying big data technologies to manage QoS in an sdn," in *Proc. 12th IEEE Int. Conf. Network and Service Management*, 2016, pp. 302–306.

[47.] Forgeat, "Data processing architectures — lambda and kappa," https:// www.ericsson.com/en/blog/2015/11/data-processing-architectures-- lambda-and-kappa, 2015.

[48.] C. E. Perkins and P. R. Calhoun, "Authentication, authorization, and accounting (AAA) registration keys for mobile IPV4," *RFC*, vol. 3957, pp. 1–27, 2005.

[49.] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Communications Magazine*, vol. 53, no. 10, pp. 190–199, 2015.

[50.] H. Zahid, T. Mahmood, A. Morshed and T. Sellis, "Big data analytics in telecommunications: literature review and architecture recommendations," in IEEE/CAA Journal of Automatica Sinica, vol. 7, no. 1, pp. 18-38, January 2020, doi: 10.1109/JAS.2019.1911795.

[51.] E. J. Khatib, R. Barco, P. Muñoz, I. De La Bandera, and I. Serrano, "Self-healing in mobile networks with big data," *IEEE Communications Magazine*, vol. 54, no. 1, pp. 114–120, 2016.

[52.] H. Isah and F. Zulkernine, "A Scalable and Robust Framework for Data Stream Ingestion," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 2900-2905, doi: 10.1109/BigData.2018.8622360.

[53.] Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower son with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[54.] R. I. Jony, A. Habib, N. Mohammed, and R. I. Rony, "Big data use case domains for telecom operators," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom*, Dec. 2015, pp. 850–855.

[55.] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with hadoop," *IEEE Network*, vol. 28, no. 4, pp. 32–39, Jul. 2014. [Online]. Available: https://doi.org/10.1109/ mnet.2014.6863129

[56.] Saad, A. R. Amran, I. W. Phillips, and A. M. Salagean, "Big data analysis on secure VoIP services," in Proc. 11th Int. Conf. Ubiquitous Information Management and Communication. ACM, pp. 5, 2017.

[57.] J. Kreps. "Kafka : a Distributed Messaging System for Log Processing." In *Proc. Kreps2011KafkaA*, 2011.

[58.] R. Van Den Dam, "Big data a sure thing for telecommunications: telecom's future in big data," in *Proc. IEEE Int. Conf. CyberEnabled Distributed Computing and Knowledge Discovery*, 2013, pp. 148–154.

[59.] Wang, J. Mi, C. Xu, Q. Zhu, L. Shu, and D.-J. Deng, "Realtime load reduction in multimedia big data for mobile internet," *ACM Trans. Multimedia Computing, Communications, and Applications*, no.

5s, pp. 1–20, Oct. 2016. [Online]. Available: https://doi.org/10.1145/ 2990473
[60.] N.R Al-Molhem, Y. Rahal, and M. Dakkak. "Social network analysis in Telecom data", *Big Data 6*, pp. 99, 2019. [Online]. Availablet: https://doi.org/10.1186/s40537-019-0264-6
[61.] Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network*, vol. 30, no. 1, pp. 44–51, 2016.
[62.] Y. Ouyang, L. Shi, A. Huet, M. M. Hu, and X. Dai, "Predicting 4g adoption with apache spark: A field experiment," in *Proc.16th Int. Symp. Communications and Information Technologies*, 2016, pp. 235- 240.
[63.] R. Siddavaatam, I. Woungang, G. Carvalho, and A. Anpalagan, "Efficient ubiquitous big data storage strategy for mobile cloud computing over hetnet," in *Proc. IEEE Global Communications Conf.*, Dec. 2016, pp. 1–6.
[64.] R. Siddavaatam, I. Woungang, G. Carvalho, and A. Anpalagan, "An efficient method for mobile big data transfer over hetnet in emerging 5G systems," in *Proc. 21st IEEE Int. Workshop on Computer Aided Modelling and Design of Communication Links and Networks*, 2016, pp. 59–64.
[65.] J. van der Lande, "The future of big data analytics in the telecoms industry," *White Paper*, 2014.
[66.] Ö. F. Çelebi, E. Zeydan, O. F. Kurt, O. Dedeoglu, Ö. Ileri, B. AykutSungur, A. Akan, and S. Ergüt, "On use of big data for enhancing network coverage analysis," *ICT*, pp. 1–5,2013.
[67.] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, "Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study," *IEEE Network*, vol. 30, no. 2, pp. 54–61, 2016.
[68.] R. K. Lomotey and R. Deters, "Management of mobile data in a crop field," in *Proc. IEEE Int. Conf. Mobile Services*, 2014, pp. 100–107.
[69.] C.-M. Chen, "Use cases and challenges in telecom big data analytics," APSIPA Trans. Signal and Information Processing, vol. 5, pp. 12, 2016.
[70.] R. Siddavaatam, I. Woungang, G. Carvalho, and A. Anpalagan, "Efficient ubiquitous big data storage strategy for mobile cloud computing over hetnet," in *Proc. IEEE Global Communications Conf.*, Dec. 2016, pp. 1–6.
[71.] Drosou, I. Kalamaras, S. Papadopoulos, and D. Tzovaras, "An enhanced graph analytics platform (gap) providing insight in big network data," *J. Innovation in Digital Ecosystems*, vol. 3, no. 2, pp. 83–97, 2016.
[72.] C.-M. Chen, "Use cases and challenges in telecom big data analytics," APSIPA Trans. Signal and Information Processing, vol. 5, pp. 12, 2016.
[73.] X. Lu, F. Su, H. Liu, W. Chen, and X. Cheng, "A unified OLAP/OLTP big data processing framework in telecom industry," in *Proc. 16th IEEE Int. Symp. Communications and Information Technologies*, Sept. 2016, pp. 290–295.
[74.] S. B. Elagib, A.-H. A. Hashim, and R. Olanrewaju, "CDR analysis using big data technology," in *Proc. IEEE Int. Conf. Computing, Control, Networking, Electronics and Embedded Systems Engineering*, 2015, pp. 467–471.
[75.] Z. Sheng, S. Pfersich, A. Eldridge, J. Zhou, D. Tian, and V. C. M. Leung, "Wireless acoustic sensor networks and edge computing for rapid acoustic monitoring," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 1, pp. 64–74, 2019.
[76.] S. Parwez, D. Rawat, and M. Garuba, "Big data analytics for user activity analysis and user anomaly detection in mobile wireless network," *IEEE Trans. Industrial Informatics*, 2017.
[77.] S. Parwez, D. Rawat, and M. Garuba, "Big data analytics for user activity analysis and user anomaly detection in mobile wireless network," *IEEE Trans. Industrial Informatics*, 2017.