ANALYTICS USING ANONYMIZED PATIENT LEVEL DATA

PROJECT REPORT

Submitted in partial fulfillment of the requirements
for the degree of

Bachelor of Technology
in
Computer Science & Engineering

Submitted b*y*

Udhav Anand (1605081)

Under the Guidance of

Prof. Anil Kumar Swain

School of Computer Engineering
Kalinga Institute of Industrial Technology
Deemed to be University, Bhubaneswar

April,2020

A PROJECT REPORT
ON

ANALYTICS USING ANONYMIZED PATIENT LEVEL DATA

(Internship at ZS Associates)

Submitted in partial fulfillment of the requirements for the award of the Degree of

Bachelor of Technology

In

Computer Science & Engineering

Submitted by

Udhav Anand
(1605081)

Under the Guidance of

Prof. Anil Kumar Swain

School of Computer Engineering
Kalinga Institute of Industrial Technology
Deemed to be University, Bhubaneswar
April, 2020

School of Computer Engineering

Kalinga Institute of Industrial Technology
Deemed to be University, Bhubaneswar

CERTIFICATE

*This is to certify that the project report entitled **"Analytics using Anonymized Patient Level Data"** has been carried out by **Udhav Anand (1605081)** in partial fulfilment of the award of degree of **Bachelor of Technology in Computer Science & Engineering** from **School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, Bhubaneswar** during the academic year 2019-2020 under my supervision.*

Mr. Shivam Chaturvedi
Decision Analytics Manager

Prof. Anil Kumar Swain
School of Computer Engineering, KIIT

Examiner 1

Date: April, 2020

Place: KIIT-DU, Bhubaneswar

Examiner 2

DECLARATION

I, ***Udhav Anand (1605081)*** hereby declare that the matter embodied in this project report is original and has not been submitted for the award of any other degree to any other university.

Udhav Anand(1605081)

ACKNOWLEDGEMENT

I would like to thank Prof. Anil Kumar Swain, School of Computer Engineering, KIIT - Deemed to be University, Bhubaneswar for his continued drive for better quality in everything that allowed us to carry out this project work.

I would also like to take the opportunity to thank the School of Computer Engineering, KIIT- Deemed to be University, Bhubaneswar for providing us with the resources and the help and co-operation in this project work.

I would like to express my gratitude to Mr. Shivam Chaturvedi, Decision Analytics Manager at ZS Associates and Mr. Siddharth G Gupta, Decision Analytics Associate Consultant and entire APLD Team for their constant support and guidance during the project.

Lastly, I would like to express our gratitude to my Parents and all the Professors, Lecturers, and friends for their co-operation, constructive criticism and valuable suggestions during the preparation and successful completion of this project report.

Udhav Anand (1605081)

ABSTRACT

Data analytics are significant apparatuses for pharma marketers, permitting them to tackle the intensity of both customary and real-world data. Utilizing data analytics through the business life cycle can carry important insights to endure that empower better targeting and comprehension of the present consumers.

Inside the pharmaceutical business, marketing and sales capacities are utilizing data analytics regularly and have been for over 10 years. In the present enormous data condition, be that as it may, there are much more ways for pharma organizations to utilize analytics in marketing and sales. For instance, organizations can pinpoint patient socioeconomics and doctor conduct through real-world data sets.

Marketers have approached data sets for a considerable length of time, which have energized traditional applications, including market sizing, patient journey analysis, pricing strategies, customer segmentations, and marketing mix.

Today, the vigor of and access to data sets is prodding creative analytic applications. Marketers have utilized data sets — especially real-world data, for example, claims information — for a long time over a huge number of utilizations. The more customary ones incorporate market sizing, patient journey analysis, pricing strategy, customer segmentation, and marketing mix, among others. New zones, for example, tolerant finding for hard-to-diagnose diseases and influence-based marketing to improve effectiveness are picking up progress.

Contents

Declaration……………………………………………………………………... ii

Acknowledgments…………………………………………………………………… iii

Abstract …………..………………………………………………………………… iv

1 **Introduction to ZS Associates ...…………………………………………………** **1**

1.1 Company Overview..…………………………………………..…………… 1

1.2 History of ZS at a Glance…………………………………………………..…. 2

1.3 What does ZS do?………………………………………………………………. 2

1.4 How ZS do, What they do?…………………………………………………… 3

2 **Literature Survey……………………………………………………………….** **4**

.

3 **Key Terminologies & About the Data set………………………………………..** **5**

3.1 Key Terminologies 5

3.2 About the Data set **6**

4 **Project Overview……………………..…………………………………………..** **7**

4.1 Datasets in the Pharmaceutical World ……………………………………… 7

4.2 Walkthrough the Project …………………………………………………….. 11

5 **Problem Statement……………………..………………………………………** **13**

5.1. Problem Definition 13

6 **Requirement Analysis…………………………………………………………..** **14**

6.1 Hardware Requirements……………………………………………………. 14

6.2 Software Requirements…………………………………………………….. 14

6.3 Program/Software/Modules Used………………………………………….. 15

6.3.1 An Analytics Workbench Tool……………………………………… 16

6.3.1.1 Hive………………………………………………………… 16

6.3.1.2 Impala……….…………………………………………… 16

6.3.1.3 Python…………………………………………………….. 17

6.3.1.4 R.……………………………………………………… 17

6.3.2 Alation……………………………………..……………… 18

6.3.3 Tableau……………………………………………………… 19

6.4. Comparitive Analysis of Technologies Used……………………………….. 20

6.4.1 Python V/s R 21

7 Project Planning……………….....………………………………………….. 23

7.1 Approach to Conducting Longitudinal Analysis………………………….. 23

7.2 Dataset Selection Criteria………………………………………………… 24

8 Implementation………….…………….………………………………………. 26

8.1 Key Concepts and Data…………………………………………………. 26

9 Conclusion and Further Scope…………………………………………….... 33

9.1 Conclusion……………………………………………………………….. 33

9.2 Challenges and Further Scope……………………………………………….. 33

10 References…………………………………………………………………….. 34

List of Figures

Figure 1.1	ZS Logo	**1**
Figure 1.2	Timeline of ZS Associates	**1**
Figure 1.3	What ZS Does	**2**
Figure 6.1	Snippet of a Flow in the Analytics Workbench	**9**
Figure 6.2	Visualization in the analytics workbench	**10**
Figure 6.3	Codes written in flow for Analysis	**12**
Figure 6.4	Alation	**13**
Figure 6.5	Visualization in Tableau	**14**
Figure 7.1	Creation of data during Paitent Lifecycle	**15**
Figure 8.1	Filtering of patients to form a clean patient cohort	**18**
Figure 8.2	Insights drawn from Data	**19**
Figure 8.3	Example of a flow for Analysis	**20**

Chapter-1

Introduction to ZS Associates

1.1. Company Overview

ZS Associates (ZS) is a US management consulting firm headquartered in Evanston, Illinois that offers types of assistance essentially in the pharmaceutical, biotechnology, medicinal services and agribusiness businesses. ZS was established in 1983 by Prabhakant Sinha and Andris Zoltners, who cooperated as educators of showcasing at the Kellogg School of Management at Northwestern University.

Fig. 1.1. ZS Logo

The firm employs in excess of 7,000 workers in more than 25 workplaces in the America, Asia and Europe. ZS's Capability and Expertise Center is situated in enormous centers in Pune and New Delhi, India. In 2014, ZS worked with 49 of the 50 biggest drug-makers and 17 of the 20 biggest medical device makers and also serves consumer products, financial services, industrial products, telecommunications, and transportation and logistics industries.

1.2. History of ZS at a Glance

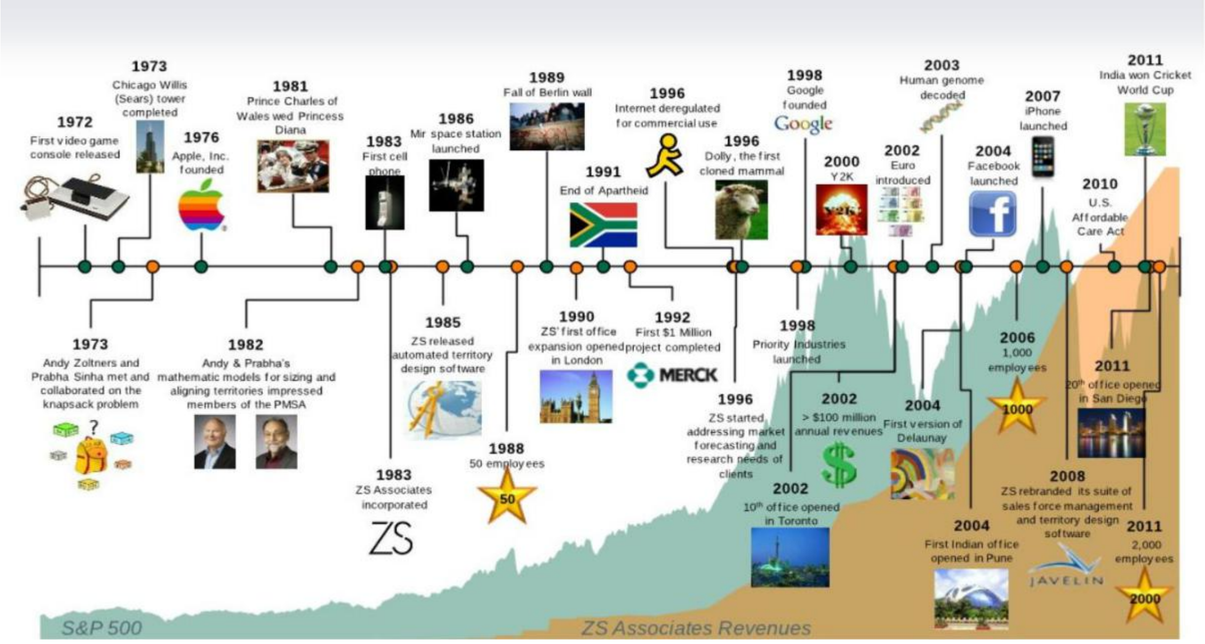


Fig 1.2: Timeline of ZS Associates

ZS was founded by Andris Zoltners, Frederic Esser Nemmers Distinguished Professor Emeritus of Marketing at the Kellogg School of Management, and Prabakant Sinha, a former associate professor of marketing at the Kellogg School of Management. At Kellogg, Sinha and Zoltners developed a side business advising companies on sales and marketing, which evolved into ZS.

As indicated by profession survey site Glassdoor, ZS Associates routinely includes among the Top 10 companies with the toughest interview process, positioning fifth in 2012 and seventh in the most recent rankings.

ZS was named one of Consulting magazine's "Best Firms to Work For" in 2017 and 2018. In light of the information gathered by Transparent Career, ZS was the twelfth best organization for MBAs to work at, while likewise being the sixth best consulting organization for MBA graduates dependent on pay, normal hours worked, and by and large employee satisfaction. A comparable report done by Wall Street Oasis dependent on the information they gathered, put ZS at the seventh spot among consulting firms having the most noteworthy job satisfaction rates.

1.3. What does ZS do?

ZS is a leading global management consulting firm. Their broadness of experience reaches out past the contributions accessible from other consulting organizations, which implies they support clients in each progression of the business procedure, from research to implementation. Be that as it may, their actual specialty – an expertise unrivaled by any other in the business – Sales and Marketing.

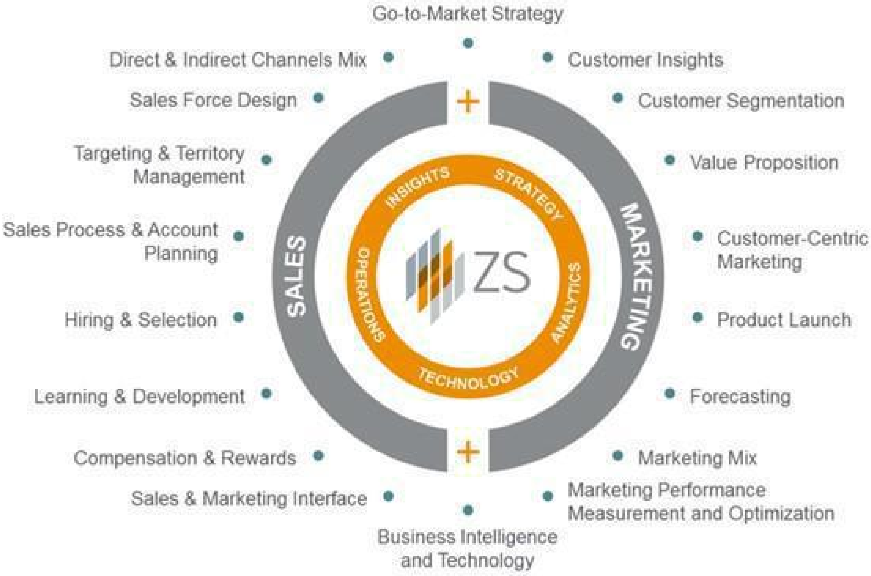


Fig 1.3: What ZS does

ZS Associates has much engaged sales and marketing expertise, wide and coordinated contributions, diagnostic and innovation abilities and industry experience. They handle ventures from customer insights and marketing techniques to product launches and technology systems – and everything in the middle.

1.4. How ZS do, what they do?

ZS don't deliver just great ideas – they stay until they're making value in reality. Their specialists work with pioneers at the world's top companies to make progressive marketing strategies and organize sales that lift piece of the overall industry i.e. Market Share, increment in revenue, lower costs and improve by and large business execution.

They ask the right questions to uncover insights that lead to better decisions. They use and fabricate the best technologies and operational procedures to plan solutions that work. And they create resolutions to hit the nail on the head the first run through.

At ZS, they keep their eyes on the future and are all set where their industry takes them. Due to the inexorably various nature of their business, they convey solutions for their clients in four areas: consulting, operations, technology and software, and integrated solutions. They additionally have an interior corporate support organization that operates ZS business functions and supports their client-facing teams.

- Business Consulting
- Business Operations
- Business Technology

Inside every one of these regions, ZSers exploit their cooperative condition to build up their ability. Since they accept the best solutions originate from pushing limits and embracing new thinking, ZSers are continually learning and developing and clients are continually getting the best.

Chapter-2

Literature Survey

In order to stay ahead of the competition, pharmaceutical and healthcare organizations need to understand the market. APLD offers companies valuable data that can help to answer important business questions about the market.

In recent years, there has been a heightened awareness that performing data analysis on individual patient data is beneficial. Tracking APLD for an extended period of time allows healthcare organizations to explore using the data intelligently, for instance, healthcare organizations can utilize evidence-based content to figure out immediate patient-specific needs as APLD is readily available compared with “aggregate” data. A good example of advanced data transformation techniques from APLD can be seen in a 2015 study that performed a meta-analysis of APLD to directly compare the differences in standard-of-care prostate cancer treatments.

Due to the use of different data sources for individual patients, healthcare organizations need to also leverage tools that have the power to standardize the data across the board. One useful emerging model for converting datasets into a standardized form is the OMOP Common Data Model (CDM). With the OMOP CDM, healthcare organizations or pharma firms can take disparate observational databases and transform them into a common format (data model), and then move forward in performing systematic analyses.

Overall, APLD is very beneficial because it is not subject to influence as clinical trials are. Therefore, APLD analytics can be leveraged to show real world performance of drugs. APLD analytics can also make comparisons to competitor drugs in real-time to influence prescription choices and formulary coverage.

Chapter-3

Key Terminologies & About the Dataset

3.1. Key Terminologies

HCP or Physician: A Doctor

Rx – Prescription written by an HCP when a patient is diagnosed with a certain disease.

Dx- Diagnosis of any disease

Therapeutic Area- An area of disease where a certain therapy is used to treat the disease.

Patient Cohort- A cleaner set of patients where all the patients share the same disease as commonality and are being treated for the same with different medication.

NRx – Rx written by the HCP when the patient visits him / her

TRx – NRx and the refills that the patient can make before revisiting the doctor

NTS: New Therapy Start – First time a patient is started on a therapy

NTB: New to Brand (NTB) – First time a patient is initiated on a brand. The patient could also be a NTS or a switch from another product in the same therapy

Copay – Out of pocket cost of the patient the fees of the doctor or the amount that needs to be given to the Pharmacy to get the prescription filled

Samples – Free samples left at the physician's office to get the new patients initiated on a drug

Vouchers – Vouchers are coupons to get free prescription for a drug from the pharmacies Copay

Cards – Cards to provide discount to the patient on the copay amount to be paid to the pharmacies to get the drugs

APLD:- Anonymized Patient Level Data, Data for Rx and Dx at a patient level where we do not know the patient name and his/her patient details but is identified by a unique Patient ID.

3.2. About the Dataset

In the pharmaceutical world, APLD has been gathered generally from exchanges among drug stores and the healthcare associations that serve and track millions of patients. These data follow anonymized patients over time with data about doctor prescriptions, drug therapies and treatment patterns.

Before the coming of APLD, organizations depended on physician level data. These data check the absolute number of prescriptions a specialist writes for every product in a specific market, yet they give little data about the patient-related factors behind a doctor's decision to recommend a specific cure.

Thus, through various data vendors, essentially, we can get a longitudinal track of each and every patient in the market. APLD contains various stages of the patient journey i.e the Rx level data, the Physician Information and their specialty, The Diagnosis information of each patient, the laboratory test results, Rx data of the patients who purchased drugs from Special Pharmacies and many more. Essentially, by doing various analysis on this data, one can track the complete journey of the patient for a treatment.

In addition to the Rx claims and Diagnosis data, we need data at physician level for various analytics and insights. Furthermore, Physician level data can be used for various other purposes.

Chapter-4
Project Overview

ZS Associates provide consultancy services to its clients in a very effective and structured way, because of which ZS is getting more and more clients day by day.

In this project, the company is dealing with a US based client which is one of the largest pharmaceutical companies by both market capitalization and sales. The client sells drugs manufactured by them to the doctors so that the doctors can prescribe the drugs to their patients. We work on analyzing the data of the client at various levels and providing them insights based on the analysis.

4.1. Datasets in Pharmaceutical World

Data sets can give insights to inform high-level strategy and enable pharmaceutical marketers to accomplish their business targets, and there are numerous sorts of data sets that hold important insights for pharma marketers. The first is patient data, which can enable a manufacturer's patient-support team break down its workload and decide case statuses. The data can inform field access teams as they work with patient-support programs, specialty pharmacies, and medical offices to help with patient access and reimbursement.

Second are transaction-level claims, such as prescription, medical, and hospital data, which assist organizations with understanding patient journeys, including the request for experts seen, diagnosis, therapies, and tests. These insights help targeting efforts and fill in as the inputs for forecasting. And third, affiliations data permit organizations to draw insights from metrics accessible at both the physician-and account-levels.

The datasets are provided to the clients by third-party data providers which basically combines the data from various pharmacies and provide the data post processing at Rx Level, HCP level and data from various special pharmacies.

Some of these datasets are as follows:-

1. **Pharmacy claims data**:- It is obtained via relationships with retail pharmacy chains and for some vendors, mail order operations. It has a broad national Rx capture; capture is typically the highest compared to all other patient-data types.

   Overview:-

   Pharmacy chains are the main source of information for pharmacy claims data

   Transactions for every product dispensed by a patient in a pharmacy is collected

   Since pharmacies are the only source, information on diagnoses, procedures or lab values are not available

   Cash transactions are also collected in the data

   Physician IDs are available (ME, DEA, NPI)

   Advantages:-

   Broad national patient, Rx, and physician capture due to open nature (i.e., patients can switch or use multiple pharmacy chains)

   Physicians are identified

   Provides payment information

   Claims paid by cash are collected in this data

   Capture of physicians, Rxs is evenly distributed

   Typical Uses:-

   Owing to its broad national capture and availability of physician ID its most useful application is physician targeting

   It can also be used to map out a patient journey to understand the buying process and the stake holders involved in every treatment decision

   Can help in understanding the source of business of a product i.e share of new Rxs, switch , add-on vs. continue.

2. **Switch Claims:-** Claims from pharmacy switches and physician office practice management software are processed and provided to payers. Switch claims are also considered an open dataset and have a robust capture across therapy areas.

Overview:-

Data is collected from physician practices and pharmacies

Information on patient diagnoses, procedures along with the dispensed Rxs are available and can be often linked at the patient level

Physician IDs are typically available

Information on rejection, reversal and approval of Rxs are available in Source Health data

Advantages:-

Broad national patient, Rx, and physician

capture Diagnoses & procedure information is

available Provides payment information

Approvals, rejections and reversals can be

obtained Physicians are identified

Vouchers and co-pay cards can be identified

Typical Uses:-

Since patient diagnoses and procedures are also captured, patient and physician segmentation is much richer

Tracking product share and usage by indication is one of most popular

uses Treatment process and source of business can also be analyzed

Promotion effectiveness can be measured as vouchers/co-pay cards can be identified

3. **Payer Claims:-** Information from commercial, Medicare and Medicaid insurance claims are captured in payer claims data. Payer claims is a fully integrated patient-data having in-patient, out-patient, diagnosis, procedure, lab information.

Overview:-

Information from commercial, Medicare and Medicaid insurance claims are captured in payer claims data

Patient capture is restricted to participating plans, hence capture rate is lower

Claims from in-patient and out-patient locations are covered in this dataset

Patient diagnoses, procedures along with the dispensed Rxs are available in this dataset

Physician IDs are not typically available

Advantages:-

Complete longitudinal capture allows for in-depth analysis of patient behaviors as a function of time, tests, and progression

Since the data is closed, interactions across the healthcare spectrum, i.e hospitals, clinics, pharmacies, labs are available

Includes Rx fulfillment and payment information

Co-morbid conditions can be identified through records of diagnoses (ICD-9 code) in physician or hospital setting

Concomitant drug therapy usage can be obtained through drug prescriptions filled at pharmacy

Typical Uses:-

Treatment process can be accurately tracked in payer claims as it has a strong longitudinal data capture

Can be used to measure health outcomes

Compliance & persistency on a medication can also be analyzed

Usage by indication can be done as patient diagnoses and procedures are also captured

4. **Hospital Claims: -** This dataset contains in-patient and out-patient medical records collected from hospitals, nursing homes and LTC facilities. Data aggregated at a hospital level can support various researches like drug utilization, cost of care, health outcomes.

Overview:-

Hospital claims data captures complete information on patients visit to a hospital (in-patient or out-patient)

Diagnosis, procedures, lab values, Medical, duration of stay, drug utilization and payment information is available in this dataset

Projection factors are provided by some vendors to roll-up data at a national level

Advantages:-

Provides rich information on individual patient healthcare utilization

The database gives a complete picture of hospitalization events, like diagnosis, procedures, lab values, medications, length of stay, medication and cost of care.

Typical Uses:-

Will be most useful for analytics in markets like oncology

Can estimate the amount of hospital spill over, an effect of promotion in the retail setting

Can be used to measure health outcomes

3.2. Walkthrough to the Project

Our client is one of the largest pharmaceutical companies by both market capitalization and sales. Thus, the client wants to analyze the data to get insights for Market Share of their products, switching trends of the patients to other drugs, counting of active writers and Rx and many more analysis that can be done by tracing the patient journey.

The analysis can be done on Anonymized Patient Level Data (APLD), upon which based on various business rules applicable for the market in that therapy area, the cohort of patients is formed. The cohort contains the cleanest set of patients that can be used for analysis so as to find the insights from the data.

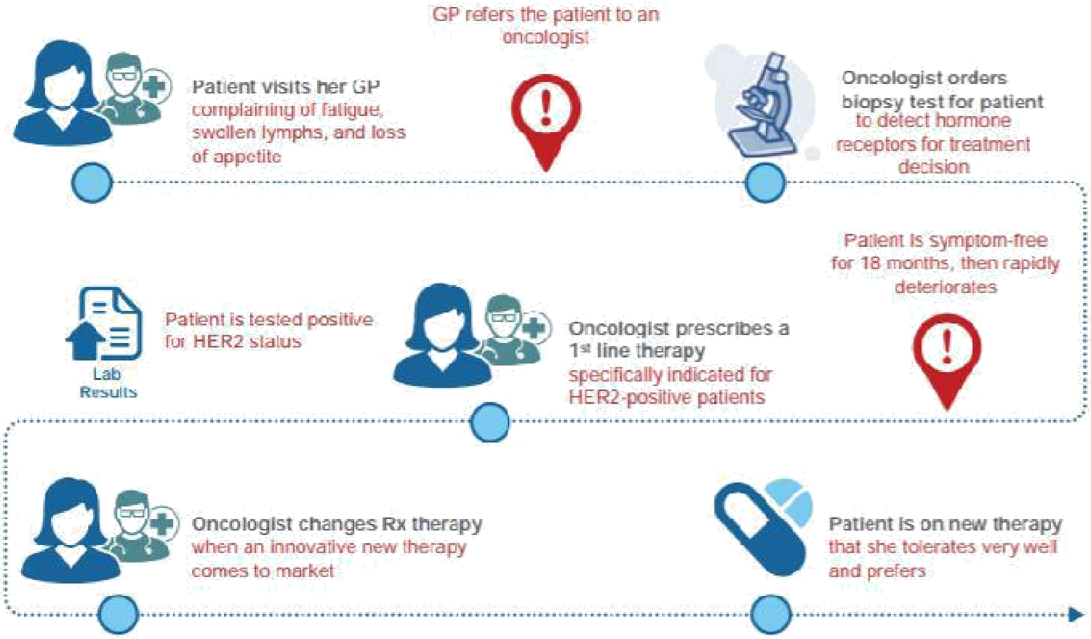

GP refers the patient to an oncologist
Patient visits her GP complaining of fatigue, swollen lymphs, and loss of appetite
Oncologist orders biopsy test for patient to detect hormone receptors for treatment decision
Patient is symptom-free for 18 months, then rapidly deteriorates
Lab Results
Patient is tested positive for HER2 status
Oncologist prescribes a 1st line therapy specifically indicated for HER2-positive patients
Oncologist changes Rx therapy when an innovative new therapy comes to market
Patient is on new therapy that she tolerates very well and prefers

Fig 4.1: Depiction of Patient Journey

Chapter-5

Problem Statement

5.1. Problem Definition

By collecting and analyzing aggregate data across the life sciences and healthcare value chains, pharma companies can derive and apply deeper insights to deliver better patient experiences and outcomes and discover new market segments for incremental revenue generation.

The aim of this project is to study the data and draw valuable insights so as to increase the revenue of the company and see the performance of the product in current scenario.

This can be done by using patient level data to identify the specific patient situations where the product can create unmatched value propositions and improve patient outcomes and draw valuable insights. Some of the situations are:

- Products with multiple indications or indications for specific patient populations.
- Products with a companion diagnostic test.
- Disease states where patient diagnosis, treatment and management involves multiple HCPs or multiple care settings.
- Disease states with complex treatment protocols, potentially involving multiple lines of therapy

Furthermore, using physician level data that provides insights in to physician behavior and decision drivers. This project lays out a strategy to help pharma companies analyze APLD data and enable business users to better understand customer segments and target their products to address the needs of those segments.

Chapter-6

Requirement Analysis

6.1 Hardware Requirements: -

- A laptop or desktop with i5 processor.
- Hard disk of 1TB.
- RAM of 16 GB.

6.2. Software Requirements: -

- 64-bit computer.
- Windows, macOS or Linux.
- Google Chrome
- SQL
- Analytics Workbench
- Alation
- Tableau

6.3. Program/Software/Modules Used: -

6.3.1. An Analytics Workbench Tool

For this project, we used a collaborative platform for all tools used in Data Science that can be leveraged by Data Analyst, Business Analyst and Data Scientist for various analytics to be performed on the dataset: -

- Import the data
- Clean, restructure and merge the input datasets together
- Split the merged dataset by whether outcomes are known and unknown, i.e., labelled and unlabeled

- Train and analyze a predictive model on the known cases
- Score the unlabeled cases using the predictive model

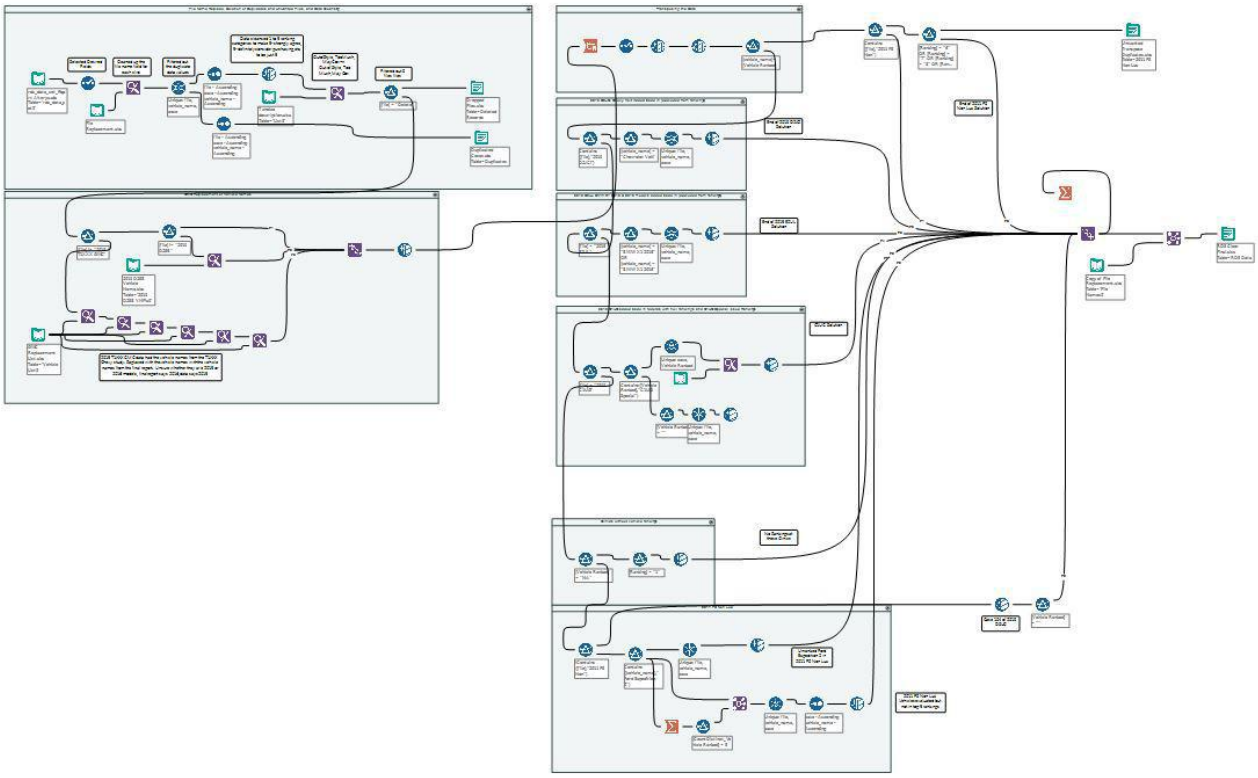


Fig 6.1: Snippet of a Flow in the Analytics Workbench

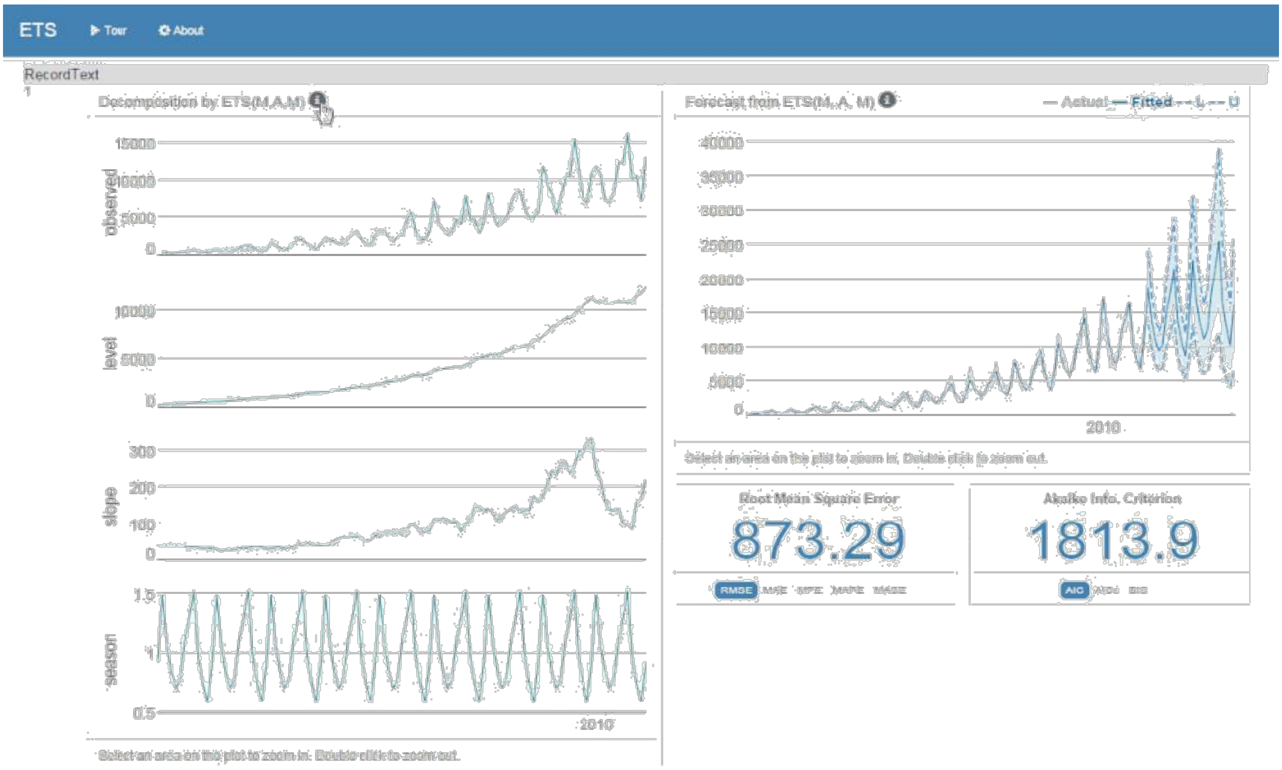


Fig 6.2: Visualization in the analytics workbench

The software allows to work on R, Python, SQL, Impala, Hive, Spark scripts inside the flow for analysis and Data Manipulation. Inside the flow, the code-based recipes are represented by red circles.

6.3.1.1. Hive

Hive is a tool of the Hadoop environment that allows running SQL queries on top of large amounts of HDFS data by leveraging the computation capabilities of Hadoop MapReduce. It can be used either as a semi-interactive SQL query interface to obtain query results, or as a batch tool to compute new datasets. Hive maps datasets to virtual SQL tables.

It provides the following integration points with Hive:

The Hive Recipe allows you to compute HDFS datasets as the results of Hive scripts

All HDFS datasets can be made available in the Hive environment, where they can be used by any Hive-capable tool, even if these datasets were not computed using a Hive recipe

The “Hive notebook” allows you to run Hive queries on any Hive database, whether they have been created by the workbench or not

6.3.1.2. Impala

Impala is a tool of the Hadoop environment to run interactive analytic SQL queries on large amounts of HDFS data. Unlike Hive, Impala does not use MapReduce but “Massive Parallel Processing”, i.e.. each node of the Hadoop cluster runs the query on its part of the data.

It provides the following integration points with Impala :

All HDFS datasets can be made available in the Impala environment, where they can be used by any Impala-capable tool.

The “Impala notebook” allows you to run Impala queries on any Impala database, whether they have been created by the workbench or not.

When performing **/visualize/index** on a HDFS dataset, you can choose to use Impala as the query execution engine.

6.3.1.3. Python

The workbench gives you the ability to write recipes using the Python language. Python recipes can read and write datasets, whatever their storage backend is.

For example, you can write a Python recipe that reads a SQL dataset and a HDFS dataset and that writes an S3 dataset. Python recipes use a specific API to read and write datasets.

Python recipes can manipulate datasets either:

Using regular Python code to iterate on the rows of the input datasets and to write the rows of the output datasets

Using Pandas data frames.

6.3.1.4. R

R is a language and environment for statistical computing. It provides an advanced integration with this environment and gives you the ability to write recipes using the R language.

R recipes, like Python recipes, can read and write datasets, whatever their storage backend is. It provide a simple API to read and write them.

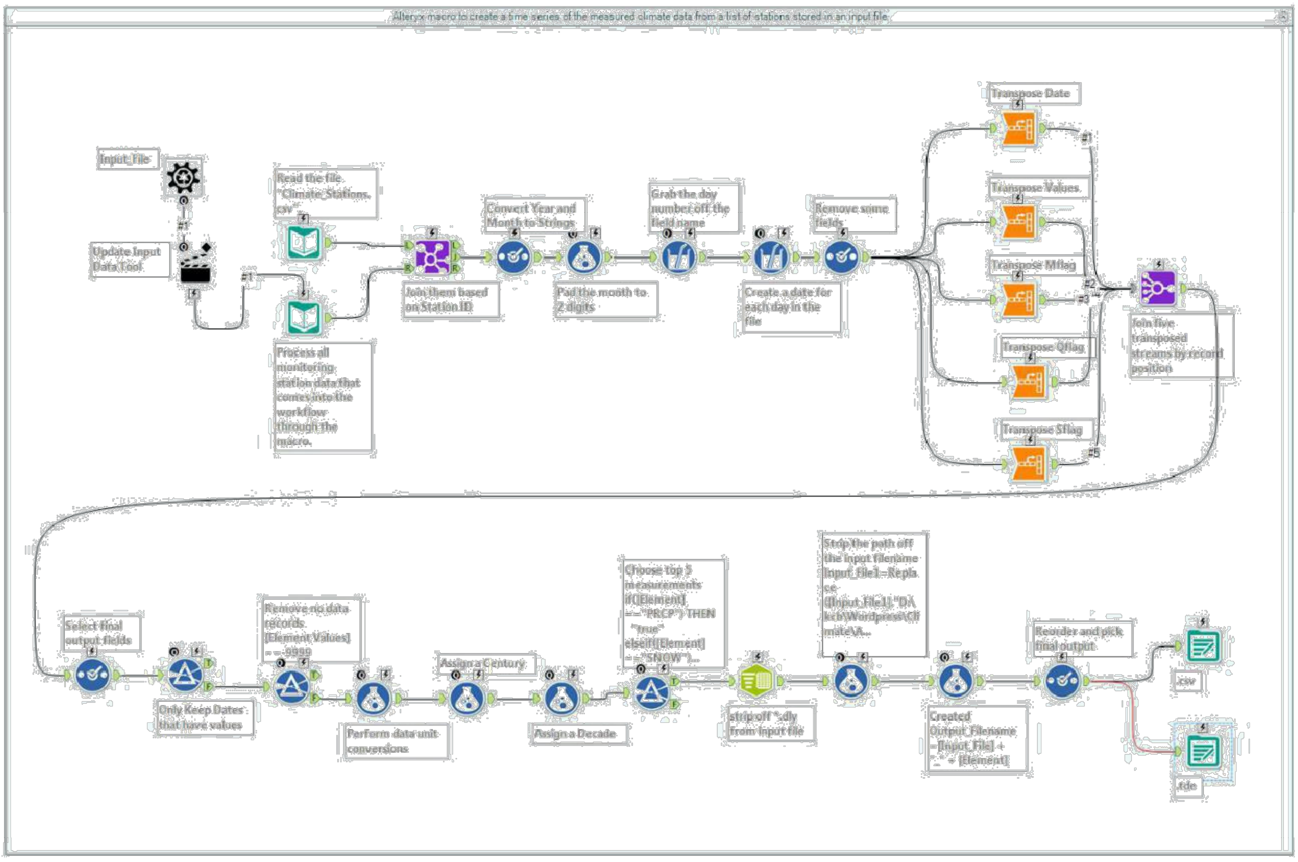

Alteryx macro to create a time series of the measured climate data from a list of stations stored in an input file
Input_File
Update Input Data Tool
#1
Read the file "Climate_Stations.csv"
Process all monitoring station data that comes into the workflow through the macro.
Join them based on Station ID
Convert Year and Month to Strings
Pad the month to 2 digits
Grab the day number off the field name
Create a date for each day in the file
Remove some fields
Transpose Date
Transpose Values
Transpose Mflag
Transpose Qflag
Transpose Sflag
#1
#2
#3
#5
Join five transposed streams by record position
Select final output fields
Only Keep Dates that have values
Remove no data records [Element Values] = -9999
Perform data unit conversions
Assign a Century
Assign a Decade
Choose top 5 measurements if([Element] =="PRCP") THEN "true" elseif([Element] =="SNOW")...
strip off *.dly from input file
Strip the path off the input filename Input_File1=Replace([Input_File1],"D:\kcb\Wordpress\Climate\A...
Created Output_Filename =[Input_File] + "_" + [Element]
Reorder and pick final output
.csv
.tde

Fig 6.2: Codes written in flow for Analysis

6.3.2. Alation

Alation is a comprehensive software for the analysts which is used to search, compose, curate and collaboarate with the data to get insights from it.

Primarily, it is used to:-

See your data

Alation is a complete repository for all the data assets & data knowledge in your organization. Alation is a single point of reference to:

Business glossary
Data dictionary
Wiki article

Understand your data

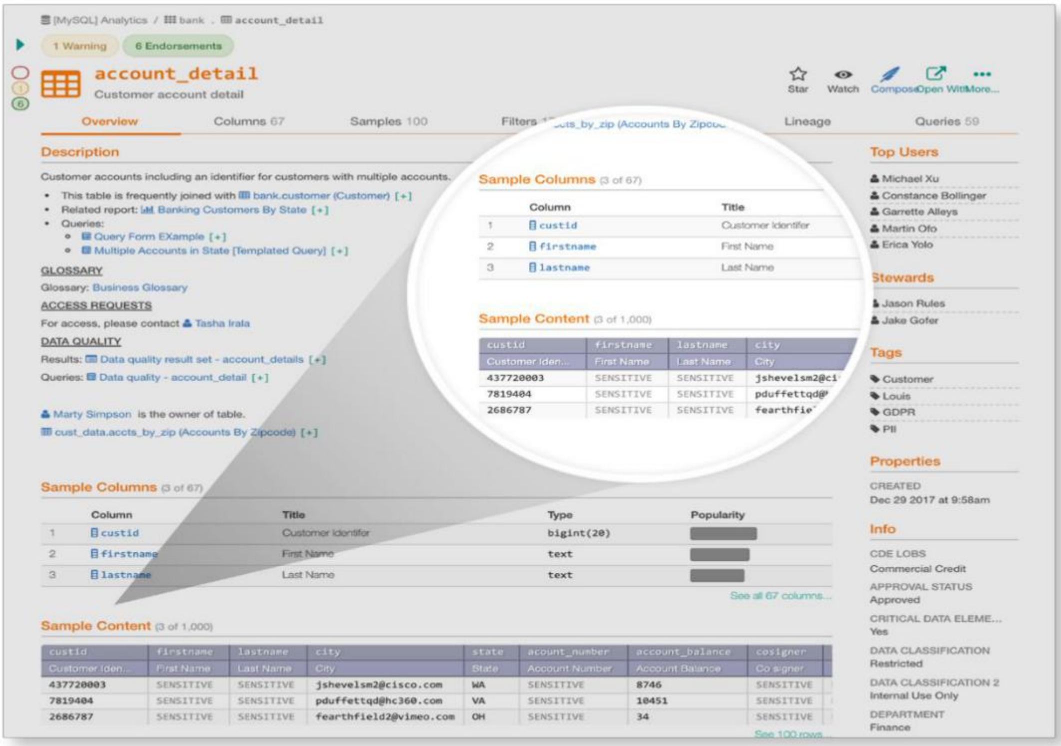
Alation samples data & monitors usage to ensure that users have accurate insight into data accuracy. This includes providing insights through:

- Catalog activity
- Data sampling
- Interactive lineage

Collaborate with data

Alation provides deep insight into how users are creating & sharing knowledge from raw data. This includes surfacing details that include:

- Top users
- Popularity of schemas, tables, and columns
- Shared joins, filters, and queries



account_detail

Customer account detail

Description

Customer accounts including an identifier for customers with multiple accounts.

- This table is frequently joined with bank.customer (Customer) [+]
- Related report: Banking Customers By State [+]
- Queries:
  - Query Form EXample [+]
  - Multiple Accounts in State [Templated Query] [+]

GLOSSARY

Glossary: Business Glossary

ACCESS REQUESTS

For access, please contact Tasha Irala

DATA QUALITY

Results: Data quality result set - account_details [+]

Queries: Data quality - account_detail [+]

Marty Simpson is the owner of table.

cust_data.accts_by_zip (Accounts By Zipcode) [+]

Sample Columns (3 of 67)

	Column	Title	Type	Popularity
1	custid	Customer Identifor	bigint(20)	
2	firstname	First Name	text	
3	lastname	Last Name	text	

See all 67 columns...

Sample Content (3 of 1,000)

custid Customer Iden...	firstname First Name	lastname Last Name	city City	state State	acount_number Account Number	account_balance Account Balance	cosigner Co signer
437720003	SENSITIVE	SENSITIVE	jshevelsm2@cisco.com	WA	SENSITIVE	8746	SENSITIVE
7819404	SENSITIVE	SENSITIVE	pduffettqd@hc360.com	VA	SENSITIVE	10451	SENSITIVE
2686787	SENSITIVE	SENSITIVE	fearthfield2@vimeo.com	OH	SENSITIVE	34	SENSITIVE

See 100 rows...

Top Users

- Michael Xu
- Constance Bollinger
- Garrette Alleys
- Martin Ofo
- Erica Yolo

Stewards

- Jason Rules
- Jake Gofer

Tags

- Customer
- Louis
- GDPR
- PII

Properties

CREATED
Dec 29 2017 at 9:58am

Info

CDE LOBS
Commercial Credit

APPROVAL STATUS
Approved

CRITICAL DATA ELEME...
Yes

DATA CLASSIFICATION
Restricted

DATA CLASSIFICATION 2
Internal Use Only

DEPARTMENT
Finance

Fig 6.4: Alation

6.3.3. Tableau

Tableau is an amazing and quickest developing data representation instrument utilized in the Business Intelligence Industry. It helps in streamlining crude data into the effectively justifiable organization.

Data analysis is quick with Tableau and the representations made are as dashboards and worksheets. The data that is made utilizing Tableau can be comprehended by proficient at any level in an association. It even permits a non-specialized client to make a tweaked dashboard.

The best features of Tableau are

Data Blending

Real time analysis

Collaboration of data

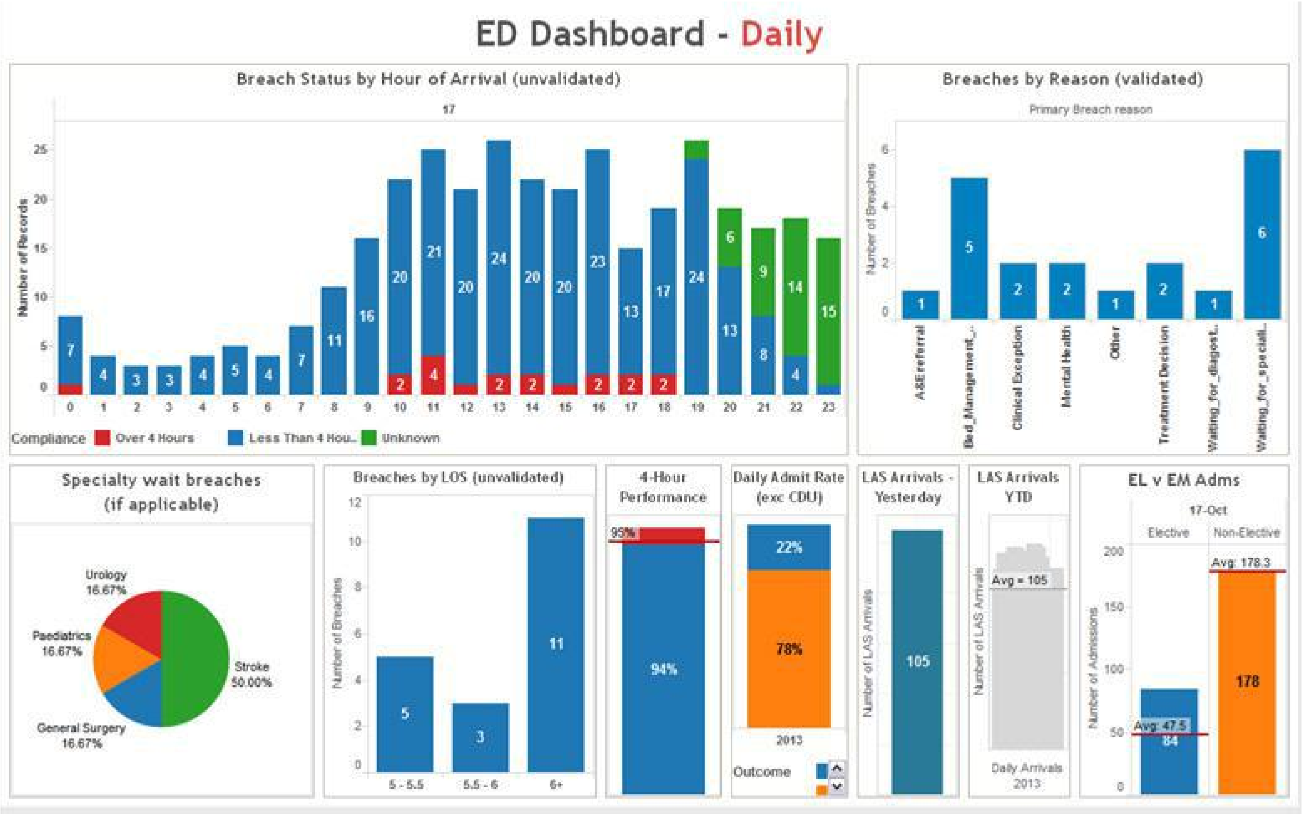


Fig 6.4: Visualization in Tableau

6.3.4. Comparative Analysis of the Technologies used:-

Hive, Impala are SQL based engines which are essentially used for Data Processing. In order to form a cleaner cohort of patients on which further analysis can be done or to find the Source of Business, finding Duration of Therapy on relational databased schema related datasets.

As far as usage of these query engines is concerned then you can consider the following points while considering or selecting any one of them:

Impala can be your best choice for any interactive BI-like workloads. As Impala queries are of lowest latency so, if you are thinking about why to choose Impala, then in order to reduce query latency you can choose Impala, especially for concurrent executions.

Hive can be also a good choice for low latency and multiuser support requirement. Do not think that why to choose Hive, just for your ETL or batch processing requirements you can choose Hive. However, Hive can reduce the time that is required for query processing, but not that much so that it can become a suitable choice for BI.

Spark SQL, users can selectively use SQL constructs to write queries for Spark pipelines. The answer of question that why to choose Spark is that Spark SQL reuses Hive meta-store and frontend, that is fully compatible with existing Hive queries, data and UDFs. Through a cost-based query optimizer, code generator and columnar storage Spark query execution speed increases.

Python and R are used for exploration of the data and once the data processing is done, to do any further statistical analysis or use the dat to perform further analysis to forecast the number or to design a machine learning model, these technologies are used.

R is mainly used when the data analysis task requires standalone computing or analysis on individual servers. It's great for exploratory work, and it's handy for almost any type of data analysis because of the huge number of packages and readily usable tests that often provide you with the necessary tools to get up and running quickly.

Python is used when the data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database. Being a fully-fledged programming language, it's a great tool to implement algorithms for production use.

6.3.4.1. R v/s Python

R	**Python**
R codes need more maintenance.	Python codes are more robust and easier to maintain.
R is more of a statistical language and, also used for graphical techniques.	Python is used as a general-purpose language for development and deployment.
R is better used for data visualization.	Python is better for deep learning.
R has hundreds of packages or ways to accomplish the same task. It has multiple packages for one task.	Python is designed on the philosophy that "there should be one and preferably only one obvious way to do it". Hence it has few main packages to accomplish the task.

R is easy to start with. It has simpler libraries and plots.	Learning python libraries can be a bit complex.
R supports only procedural programming for some functions and object-oriented programming for other functions.	Python is a multi-paradigm language. It means python supports multiple paradigms like object-oriented, structured, functional, aspect-oriented programming.
R is a command line interpreted language.	Python strives for simple syntax. It has a similarity to the English language.
R is developed for data analysis, hence it has more powerful statistical packages.	Python's statistical packages are less powerful.
R is slower than python but not much.	Python is faster.
R makes it easy to use complicated mathematical calculations and statistical tests.	Python is good for building something new from scratch. It is used for application development as well.
R is less popular but still, it has many users.	Python is more popular than R

Chapter-7

Project Planning

7.1. Approach to Conducting Longitudinal Analysis

A patient-driven methodology is advantageous to all stakeholders in the life sciences and healthcare businesses. We apply APLD in client engagements close by conventional data to give metrics that offer data at the patient level, which is increasingly granular and subsequently yields progressively exact bits of knowledge into patient and physician behavior. This would then be able to be utilized for strategic decisions that make more prominent value for physicians, patients and payers.

Essentially, APLD is data that can be mapped to individual patients, longitudinally i.e. the data can be used to track the journey of a patient throughout the treatment. The journey includes the interactions with the physicians/HCPs and can also be used to identify the patients who were diagnosed with a disease and can even used to track the list of medicines that were consumed by the patient throughout the journey. The patient-level data is collected from various components of the healthcare system (e.g., pharmacy, hospitals/clinics, payers and physicians) and compiled as a longitudinal database

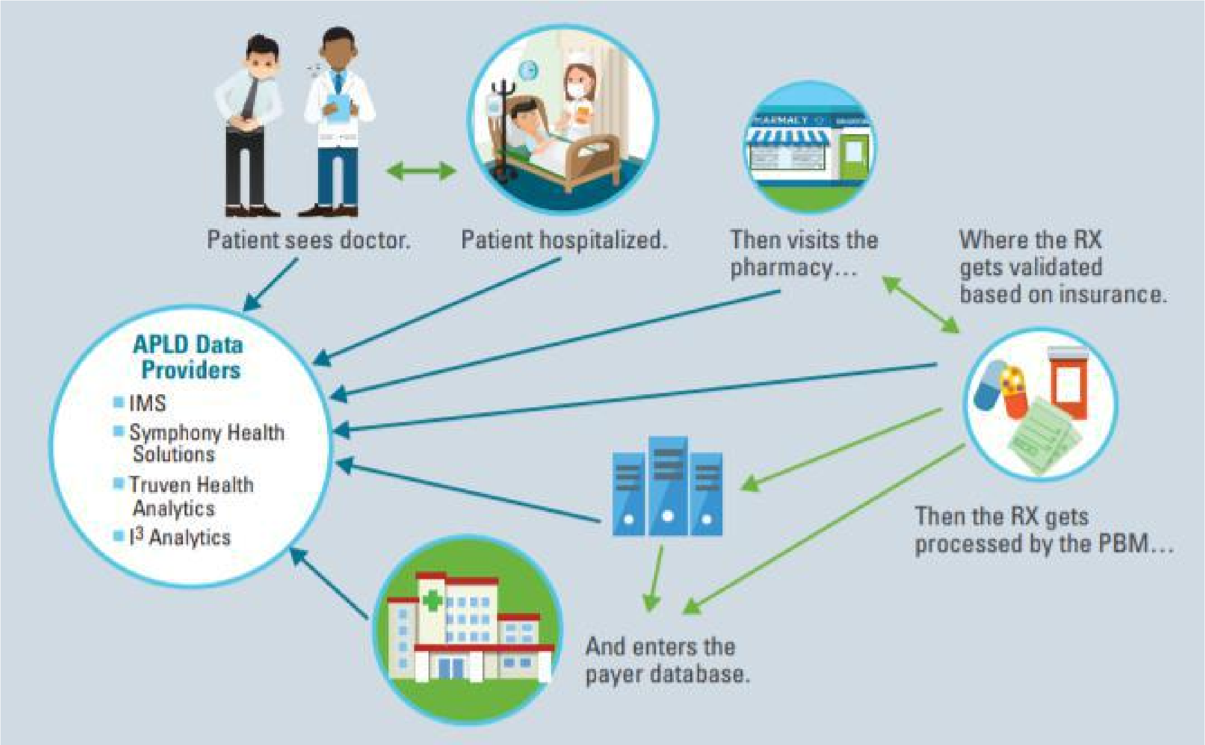

Patient sees doctor.
Patient hospitalized.
Then visits the pharmacy…
Where the RX gets validated based on insurance.
APLD Data Providers
IMS
Symphony Health Solutions
Truven Health Analytics
I3 Analytics
Then the RX gets processed by the PBM…
And enters the payer database.

Fig. 7.1. Creation of Data during Patient Lifecycle

For maintaining privacy and for preventing the data leakage it can be accessed through a server remotely. This data pertains to a therapy area and can be filtered as per disease of interest. This data can be filtered out and refined to a greater extent by defining right set of business rules which can be used to yield various insights and metrics that can be used by the clients. These business rules are defined after having an in-depth knowledge about the therapy area, the diagnosis and treatment of the disease and about other proxy drugs in the same market of the therapeutic area.

In order to get a proper universe of patients for analysis in a given therapeutic area, we can filter out the patients of our interests i.e. get a cleaner cohort based on the business rules of the therapy market and work on the cleaner cohort for any further analysis of that market to draw any insights.

Once the cohort is ready, we can work on the same to find Patients on a certain drugs on a monthly basis to get the trends, analyze the switching patterns of the drugs by the HCPs, finding the market share of the drugs with a comparative analysis of the competitors, Duration of Therapy, New Patient Start on a certain drug on a monthly basis.

The analysis can be done on physician level data as well to find the active writers' trends and we can work on other data sets to analyze the trends of detailing done by the Sales Reps so as to be used for Incentive Compensation.

7.2. Dataset Selection Criteria

The datasets are available at various levels and for different analysis we need different datasets hence the selection criteria of choosing the right dataset plays a vital role for doing the analysis accurately.

Some of the commonly used criteria are as follows: -

1. **Size**
    - Number of scripts captured
    - Patient lives
2. **Identity of Physicians**
    - Physician ID will be key for tactical project
3. **Diagnoses, Procedures and Lab values**
    - Can help in profiling patients based on co-morbidity, severity of disease etc.
    - Diagnoses can help track usage by indication
4. **Open vs. Closed**
    - Open data sources (eg. Pharmacy, Switch) does not provide reliable longitudinal medical information

- Closed data sources (eg. Payer claims) can track complete medical history of a patient. Useful for health outcomes or treatment process analysis

5. **Cash, Mail-order and specialty pharmacy capture**

    - Pharmacy claims captures cash

    - Relevance of mail order or cash is dependent on the therapy area

Some of the examples on selection of the dataset criteria based on the objective of the analysis of the project:-

1. **Targeting physicians based on their brand initiation behavior**

    a. Physician ID is necessary, High Rx capture

    b. IMS, SHS (Pharmacy/Switch claims)

2. **Track usage by indication**

    a. Require Rx and Dx information

    b. SHS, EMR, Payer claims

3. **Understand the buying process in a market**

    a. Require longitudinal medical information of patients

    b. EMR datasets are preferred as they have robust longitudinal capture

    c. SHS, IMS are open databases although with lower longitudinal capture

Chapter-8
Implementation

8.1. Key Concepts and Data

There are different kind of data on which the analysis can be performed to draw insights.

Account (DDD) Level Data tracks in-sells units and dollar sales at outlet level such as hospitals, clinics, and veteran affairs and is crucial for account-based products

Payer or Plan Level Data captures HCP prescription volume by patient health insurance plans and is a key data set in Managed Care Access related projects

Patient Level Data can also capture patient interactions with physician, hospital, and pharmacy to provide a longitudinal view of patient

Selection of the data source will be driven by the requirement gathering and elicitation.

The analysis done on Anonymized Patient Level Data (APLD) can draw various insights such as finding the source of business i.e. whether a patient is New to Brand, New to Therapy, what is the market share of the drug, how is the drug performing compared to other drugs in the therapeutic area, what is the duration of therapy of a certain drug,

Not only that, we can do analysis on Physician Level Data which can be used to draw insights on Physician/HCP prescribing behavior, finding the no. of Active Writers for a certain drug just to get an idea on how Sales Reps are performing, since the active prescription directly influence the sale of the certain product.

APLD contains the information of all the patients diagnosed or being treated by any of the drugs that has been prescribed by a physician. It contains the instances of all the instances of the interaction between Physician and Patient.

In order to do an analysis on any data to draw insights, the following steps are followed:-

1. **Requirement Elicitation** includes the research and expertise about the therapeutic area, action mechanisms of the drugs and information about the drugs which have approval for the similar indication and drugs which are approved for more than one indication.
2. **Feasibility Analysis** pertains to whether the analysis can be performed for the therapeutic area with the available data sets or if there is need of any additional data.
3. **Patient Cohort: -** The patient cohort contains the cleaner set of patients based on various rules defined with the help of requirement elicitation. This includes the use of proxy drugs approved for the indication for the same market, Patients diagnosed with the same indication,

patients being treated with the therapies specific to the therapeutic area and various laboratory test results.

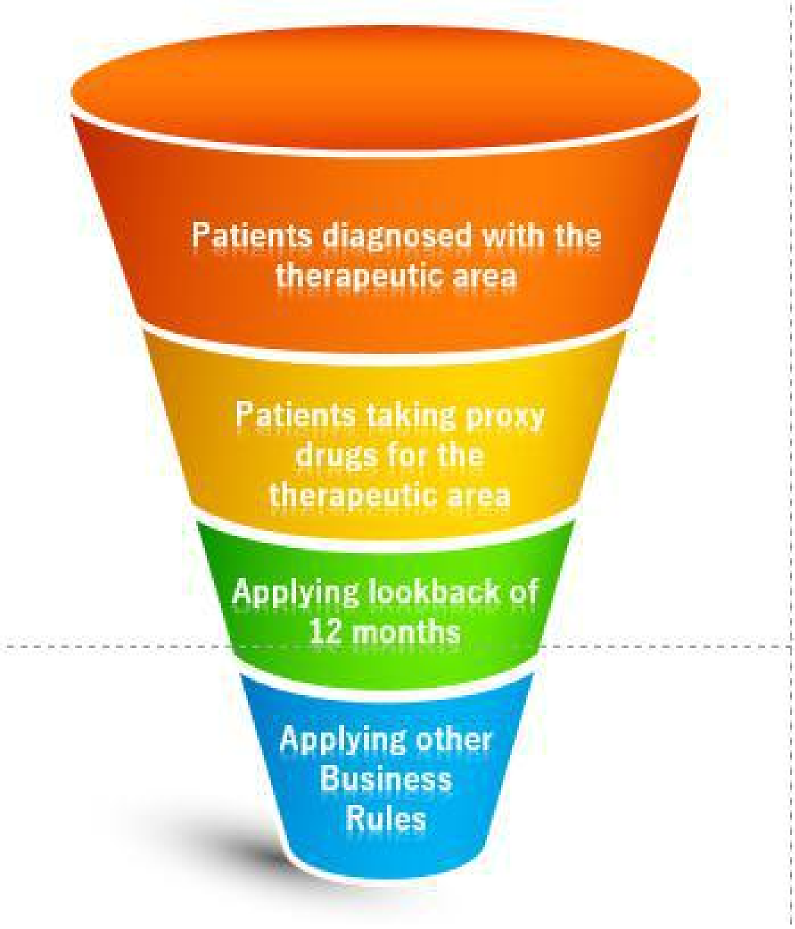

Patients diagnosed with the therapeutic area
Patients taking proxy drugs for the therapeutic area
Applying lookback of 12 months
Applying other Business Rules

Fig 7.1: Filtering of patients to form a clean patient cohort

The patient cohort is used to perform analysis to draw insights. Some of these insights include: -

- **New Patients/New to Brand/New to Market**: - Look Back is the minimum time that a patient should not be on medication to confidently classify as New to Market or New to Brand. It is an input to various analysis such as Source of Business, Line of Therapy, etc.
- **Continuing Patients**: - Grace Period is the time within which a patient should refill the prescription to be classified as continuing on drug. It is an input for metrics such as length of therapy, compliance etc.
- **Switch in Drugs: -** An Rx is defined as a switch Rx if the patient is taking a drug along with another drug in the same market. This can be calculated by: **-**

  Checking the gap between two R/x of the same product. If the gap between two R/x of same product for that patient is **greater** than the grace and the patient is taking another drug for the same market in between the aforementioned gap.

- **Add-on Drug: -** An Rx is defined as an Add-on Rx if the patient is taking a drug along with another drug in the same market. This can be calculated by: -

  Checking the gap between two R/x of the same product. If the gap between two R/x of same product for that patient is **greater** than the grace and the patient is taking another drug for the same market in between the aforementioned gap.

- **Line of Therapy: -** It is essentially a implication for change in therapy as a result of worsening of the disease. A line change in patient is defined by addition of a new drug to the previous regimen of combination of the drug therapy or switch to another drug to the previous regimen of combination of the drug therapy

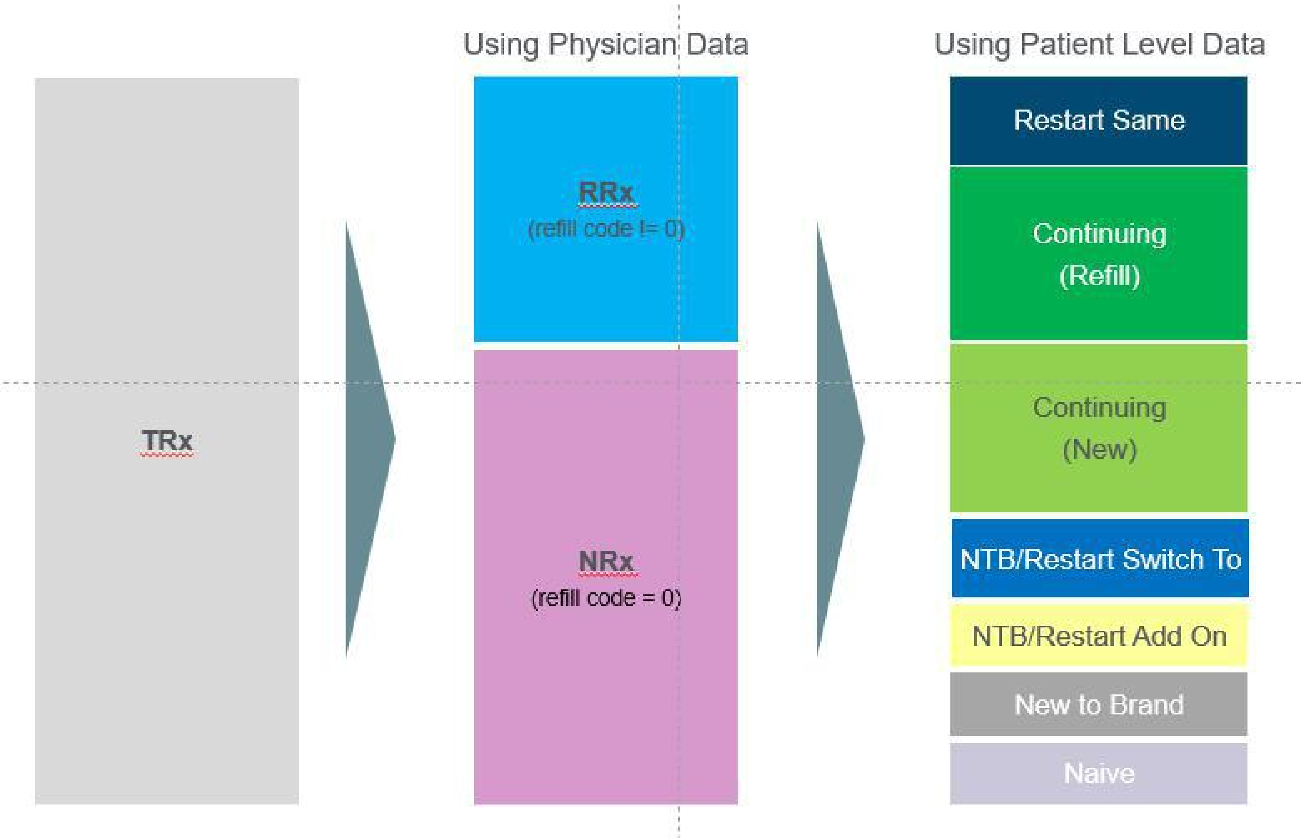


Fig 8.2: Segregation of Patients for Source of Business

- Furthermore, there are various other insights which can be drawn from the data such as Duration of Therapy by calculating the average at episode level, patient count analysis.
- From other data sources, analysis can be done to calculate the prescribing behavior of the HCP, Target Customer List analysis for pitching to HCPs to increase the sales of the product as per different market group, new and continue HCP prescribing trend analysis, TRx count analysis for various product in the market, market share of the products in a specific market and many more.

All these insights can be drawn from the data by building a flow in the analytics workbench with the help of various code-based recipes such as Python, SQL, Impala, Hive etc. and other in-built recipes.

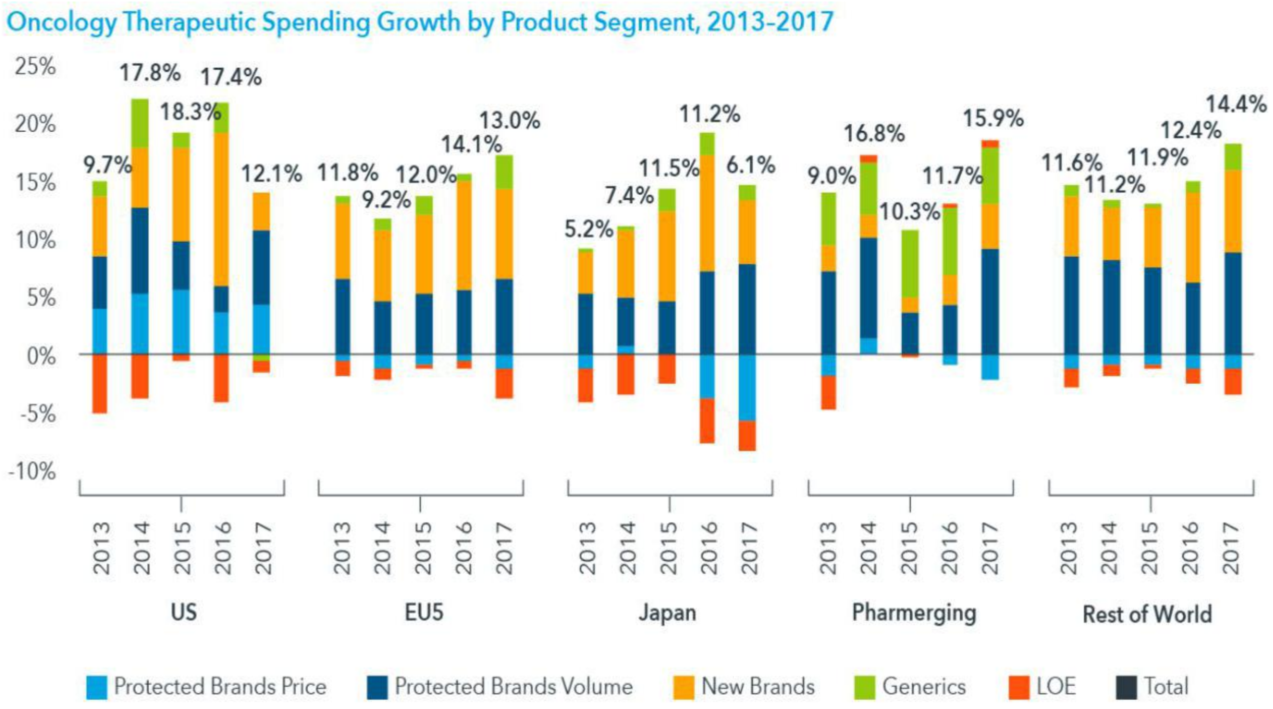


Fig 8.3: Insights drawn from Data

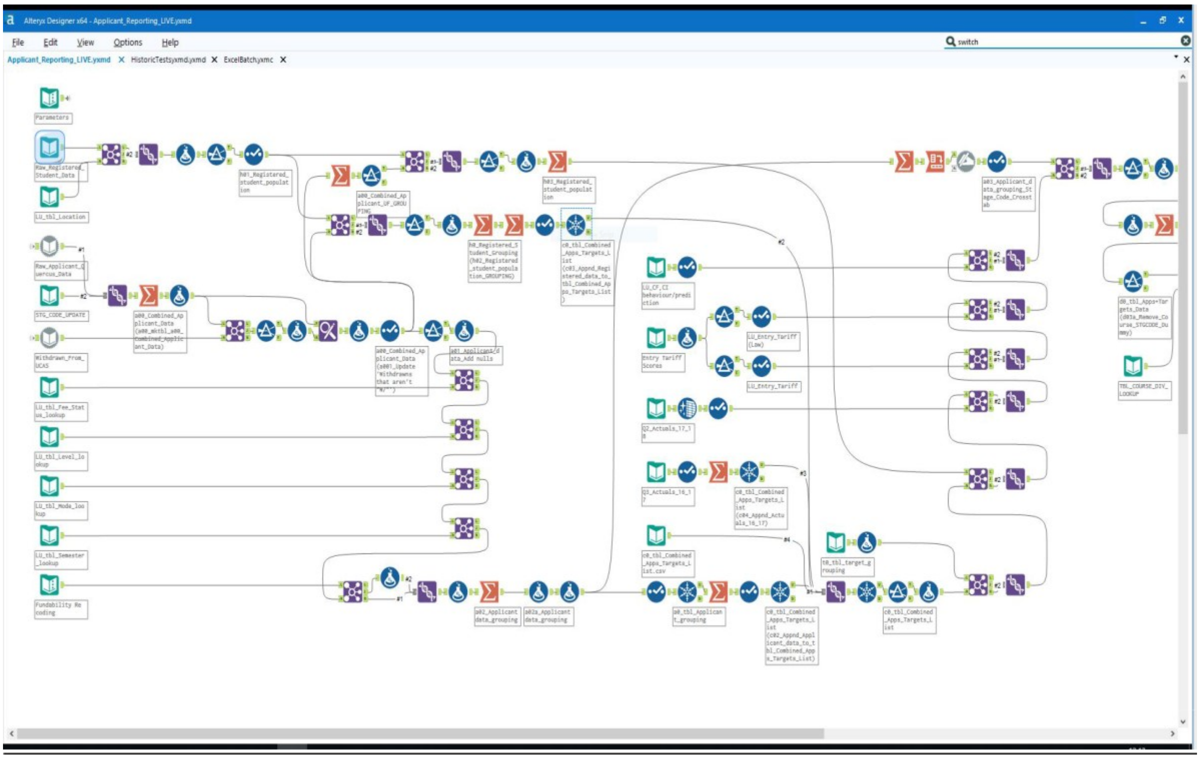

Alteryx Designer x64 - Applicant_Reporting_LIVE.yxmd
File
Edit
View
Options
Help
switch
Applicant_Reporting_LIVE.yxmd
HistoricTestsyxmd.yxmd
ExcelBatch.yxmc

Fig 8.4: Example of a flow for Analysis

Mentioned below is one of the enhancements that can give an overall idea about the use of analytics to draw meaningful insights from the data.

Scenario:

The client is a major pharmaceutical company and ZS was approached to help with one of the Oncology markets for which the client manufactures a drug. There are various other competitor firms which manufactures drug for the same indication, these drugs include drugs which were approved for only one indication and drugs that have been approved for more than one indication.

Based on the steps mentioned above we come up with a cleaner cohort for the market on which the analysis is being done. This leads to creation of the flow for drawing various market insights while carrying out analysis.

Now, since there were various other competitors drugs, the client mentioned the approval of another competitor drug in the market which needs to be included in our analysis for keeping track of the market share to see the performance of our drug and various competitors in the market and to keep an eye on the prescribing behavior of the HCP and the positioning of our drug during detailing by our sales reps.

Approach:

ZS solved the problem with the following approach:-

1. We started with researching about the new drug (hereby mentioned as Drug A) that was approved by the national drug agency in the market or therapeutic area where our client's drug is being used for the treatment. Drug A was approved for multiple indication one of which had no relation pertaining to the therapeutic area of which the analysis is being done.

2. Now, this led to further dig into research about Indication P and Indication Q (both are the indication for which Drug A is approved.) The research was primarily done with an objective to come up with a basis to identify the patients who are taking Drug A for indication P since that falls into the targeted therapeutic area.

3. The research led to some laboratory tests which can help us differentiate between patients taking drugs for the two-given indication. However, due to non-availability of the test results in the current datasets, we had to research more. This led us to find about a Drug B which is a competitor of Drug A for Indication Q (the non-targeted therapeutic area).

4. Drug B was approved by the National Drug Agency just for one indication i.e. Q. This acted as a filter to find a cleaner cohort of patients taking Drug A for targeted Indication i.e. P. Essentially because if we see a patient has a history of taking Drug B, this means that the patient has switched from Drug B to Drug A, as prescribed by the HCP for the treatment of Indication Q.

5. However, we are still unsure if the rest of the patients are taking Drug A for Indication P since we see a lot of data discrepancies and data capture issues at various stages of the Patient Journey. Thus, before adding the filtered patient pool into the actual Patient cohort, we need to ultra-refine to make sure we are not including the doubtful patients for Indication P in our analysis since that will lead to biasing.

6. Thus, taking data capture issues into account, we can take those patients who are either diagnosed with disease (Indication P) since those patients would be essentially be taking Drug A for indication P since they have been diagnosed for the targeted indication. However out of the non-diagnosed patients, we can further find the patients who have taken any other Indication P drug (drugs that are currently used in the analysis to form patient cohort) for the targeted therapeutic area, because those patients would essentially be taking Drug B for indication P since they have a history of taking drug for the targeted therapeutic ares.

7. However, we are unsure about the later patients and would not include those patients into the existing Patient cohort.

8. Once the filtered patients are added to existing patient cohort, we would be applying business rules of lookback (if any) and essentially creating episodes and regimens to come up with the line of therapy.

9. Line of Therapy can help us to know if the drug is prescribed to a patient in the starting of his journey or at a later point since the line changes on a addition of a new drug or switch to a new drug. This would be beneficial to know the prescribing nature of HCP and would even lead to increase in sales.
10. The new cohort can now be used to draw meaningful insights as and when needed. This lays out a strategy to help the client analyze APLD data and enable business users to better understand customer segments and target their products to address the needs of those segments.

Chapter-9

Conclusion and Future Scope

9.1. Conclusion

Confronted with the difficulties of the 21st century, the present pharmaceutical furthermore, biotech companies need the most ideal market intelligence they can command. Appropriately deciphered and applied, Anonymized Patient Level Data fills the requirement for more honed customer focus. The significantly improved insights by patient-level data can help guarantee better utilization of promotion dollars, progressively effective launch of the product, increasingly profitable sales manager sales rep and sales rep-HCP, and the more obviously graphed course toward improved sales rep-doctor relations and gainful growth.

9.2. Challenges and Future Scope

While the conclusions we can make from thoroughly examined APLD are clear, analyzing the voluminous and unwieldy crude data is inalienably entangled. Patient level data originate from numerous sources as opposed to a census; and each source has its own qualities and impediments. State laws overseeing medicine tops off connected to HCP visits can convolute new medicine and patient-related brand decisions made by a practitioner.

The three sorts of patient — New, Continuing and Switch — are connected after some time for an individual patient, so they can't be concentrated in seclusion. Confusing issues further, data vendors have various ways to deal with classification and identification patients

Chapter-10
References

- https://www.zs.com/-/media/files/publications/public/whitepapers/zs-thinking-beyond-the-average-patient.pdf?la=en
- http://www.fiercebigdata.com/story/obamacare-spurs-growth-data-analytics/2013-09-10 2
- http://www.todaysgeriatricmedicine.com/archive/0115p12.shtml 3
- https://www.digitalnewsasia.com/insights/four-top-trends-in-healthcare-data-analysis 4
- http://www.pm360online.com/new-ways-to-evaluate-physicians/

MAJOR PROJECT REPORT

ORIGINALITY REPORT

7%	7%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	doc.dataiku.com Internet Source	4%
2	www.cognizant.com Internet Source	2%
3	en.wikipedia.org Internet Source	1%
4	www.ukessays.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 12 words
Exclude bibliography	On		