PRIME INTUIT

Finishing School

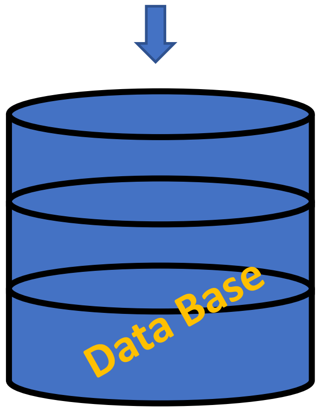
Stages of Data Science

Name:

Mobile # : 9XXXXXX45
Email ID : xyz@gmail.com
DOB/Age   : 01/01/2000
Login Name: XYZ
Password : *******
Card Details:2015 ** *****

Delivery address:

853, 1st Main Road,
6th cross, JP Nagar
3rd phase, Bangalore
560078



DBMS

What is Data:

Facts or Statistics about people, person or object
Qualitative or Quantitative variable

RDBMS

Address Table

Address #	House #	Street name	Cross	Main	City
101	853	Kothnur	1st	2nd main	Bangalore

Customer Table

Customer ID	Name	Mobile	Email ID	DOB	Log in	Card	Address
12211		9xxxxxx45	xyz@gmail.com	01/01/00	xyz	*****	101

Order Table

Customer ID	Order #	Part #	Desc	Qty	UoM	Price	Tax %	Total
12211	1100	M2231	iPhone 13	1	EA	75000	18	88550

Delivery Table

Order #	Delivery #	Part #	Desc	Qty	UoM	Address	Freight
1100	2001	M2231	iPhone 13	1	EA	101	DHL

Stages in Data Science

Collect

Store

Process

Describe

Model

Collecting Data

Flipkart

Data Already Exist

Skills

python

Java

SQL

Public Opinion

Data Already Exist

in f You Tube

Bē

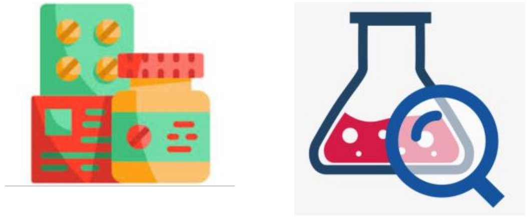
Skills

API

Java

python

Pharma & Chemical



Data Doesn't Exist

Design Experiments

Storing Data

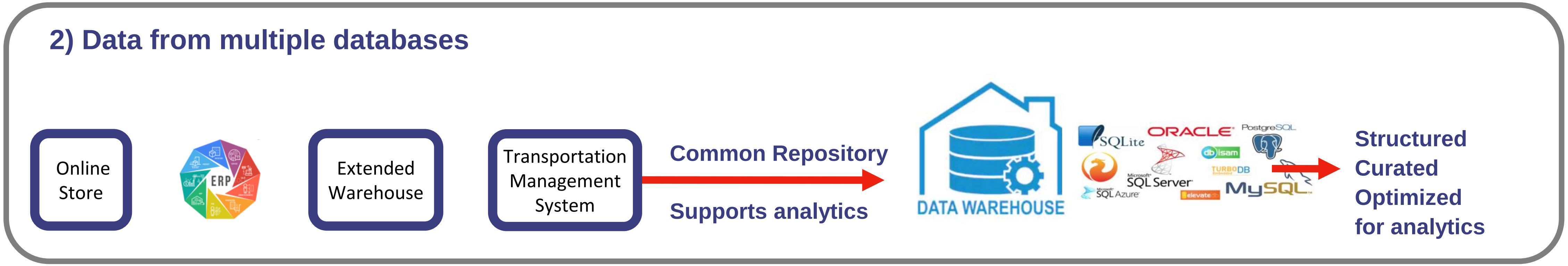
1) Master and Transactional data

Cust ID	Name	Address	Account
20001	Ravi	Mysore	1200001
20002	Sunil	Bangalore	1200002
20003	Hari	Bangalore	1200003
20004	Ganesh	Mangalore	1200004

RDBMS →

Structured
Optimized for SQL

Storing Data



2) Data from multiple databases

Online Store

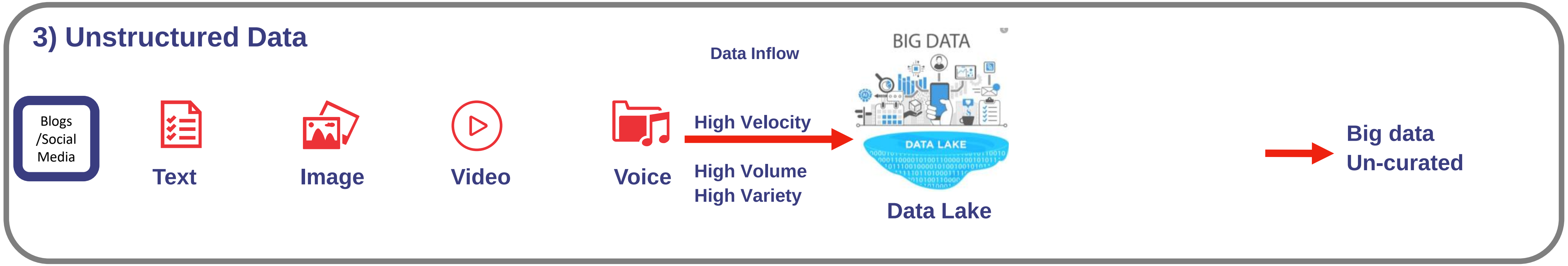
Extended Warehouse

Transportation Management System

Common Repository

Supports analytics

Structured
Curated
Optimized
for analytics



3) Unstructured Data

Blogs /Social Media

Text

Image

Video

Voice

Data Inflow

High Velocity

High Volume
High Variety

Data Lake

Big data
Un-curated

Processing Data

1) Data Wrangling and Data Munging

Part #	Time Stamp	First Name	Last Name
120011	3012020193000	Ravi	Kiran
120012	9112020193000	Sunil	Pawar
120013	8112020193000	Hari	Prasad

```
{
Item_name: “XF120013”
delivery_date: “8 Nov 2020”
delivery_time: “19:30:00”
Customer: “Hari Prasad”
}
```

2) Data Cleaning

- ○ Fill missing values
- ○ Standardize keywords tags
- ○ Correct spelling errors
- ○ Identify and remove outliner

3) Data Scaling, normalising, standardising

○ Scale ○ Normalise ○ Standardise

Skills Required:

- Programming Skills
- Map Reduce (Hadoop)
- SQL and NoSQL Databases
- Basic Statistics

Describing Data

1) Visualising Data

2) Summarising Data

○ Mean ○ Median
○ Mode ○ Variance
○ Std deviation

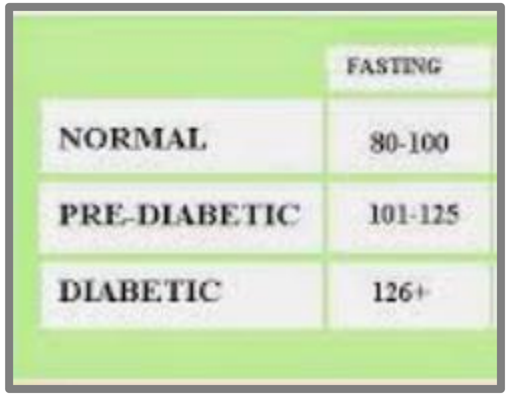
○ Descriptive Statistics
○ Iterative Process
○ Exploratory Data Analysis

Skills Required:

- Statistics
- Excel
- Python
- Tableau

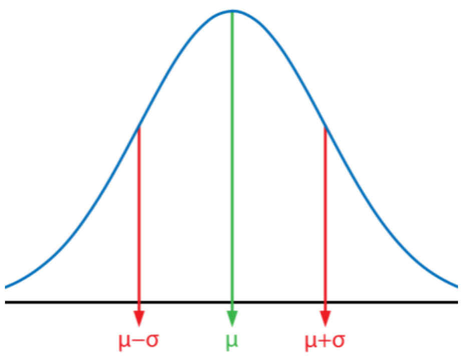
Modelling Data

Statistical Modelling



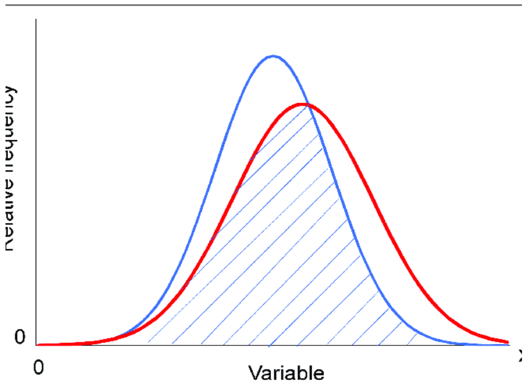
	FASTING
NORMAL	80-100
PRE-DIABETIC	101-125
DIABETIC	126+

1) Underlying data distribution



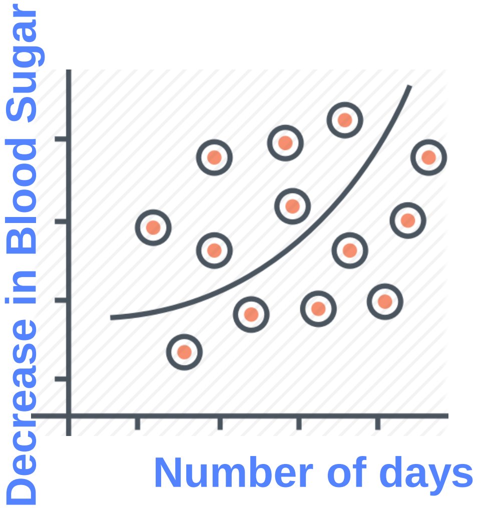
74
78
80
89
94
97
104
113
119
124
129
131
142

2) Underlying relations in data



74	74
80	78
87	80
91	89
95	94
111	97
112	104
115	113
119	119
123	124
126	129
130	131
137	142

3) Give Statistical Guarantees



Y = mx+c

Application:

- Simple Models
- Allows robust statistical analysis
- Gives Statistical guarantee results

Modelling Data

Algorithmic Modelling

1) Focus on prediction and not the phenomena

Y = f(X)

↓

(Age, weight, blood pressure, Hight, gender……….)

Y = m1x1 + m2x2 + m3x3 + m4x4 + m5x5…………..mnxn

Y = f (x1,x2,x3,x4,x5……..xn)

Here we can estimate value of f using data, optimization technique

For new patients plug input the value of x to get y

Skills Required:

- Inferential Statistics
- Probability Theory
- Calculus
- Optimization Algorithms
- ML & DL
- Python Packages and frame work (numpy, scipy, scikit-learn,PyTorch etc.