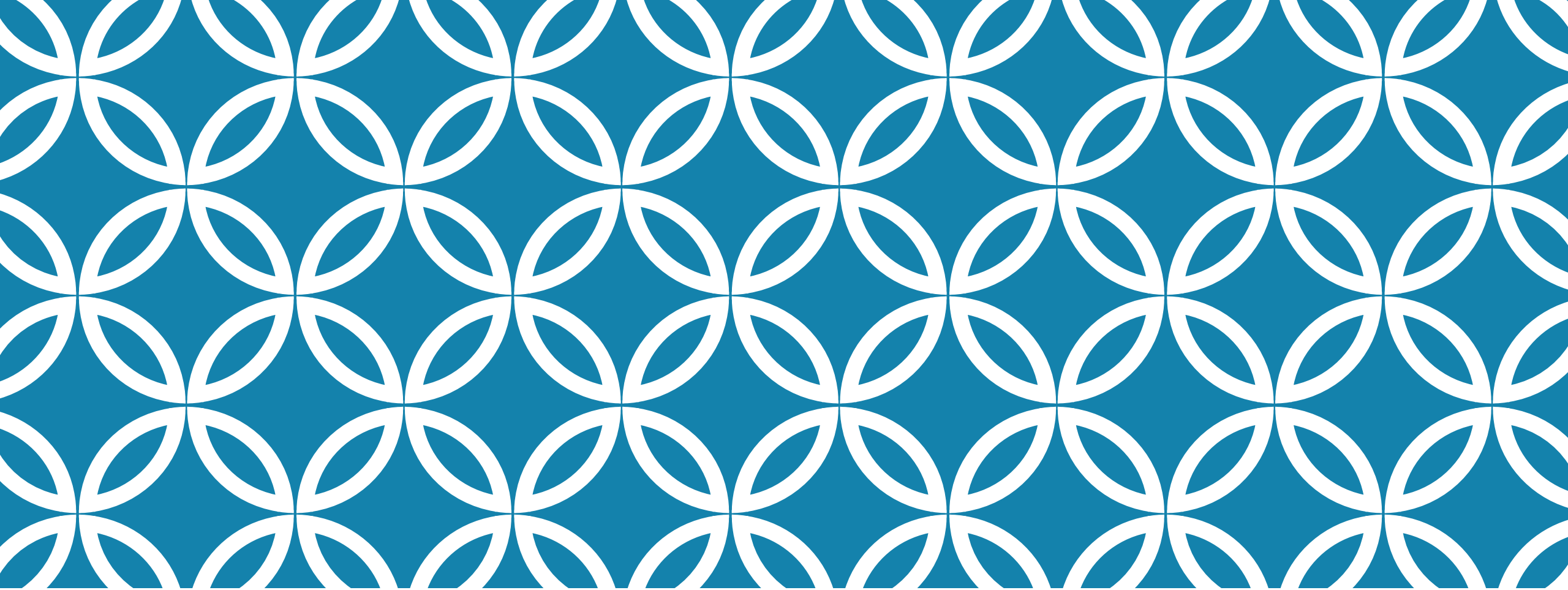


ACCELERATING ANALYTICS WITH DATABRICKS AND AWS S3

Krishna Sunder (2261141)

PROBLEM STATEMENT

K-Mart, is an Indian retail corporation that operates a chain of hypermarkets which is present in 10 locations across India. They have incurred heavy losses in the past 2 years due to flawed decision making. To become profitable, the company needs assistance in:

i. Setting up a connection between AWS S3 and data bricks and upload historical normalized sales data into S3 bucket.

ii. Identifying products with maximum and minimum demand in a particular city, year, month and subcategory

iii. Identifying the city that has generated the least revenue over the two years, so that a store can be opened in a city producing maximum revenue, and a store can be closed in the city producing minimum revenue.

TEAM

AJAY SIVARAMAN

SCRUM MASTER

KRISHNA SUNDER

DATA ENGINEER

SARVESH MISKIN

DATA ANALYST

ANISHA KALE

DATA ANALYST

SWATI THAKRE

PRODUCT OWNER

ARAVIND ROYAL

DATA ENGINEER

PROCESS FOLLOWED

Data Cleaning

1

Normalization

2

Linking Data Bricks to S3

3

PROCESS FOLLOWED

Uploading Normalized Data to S3

4

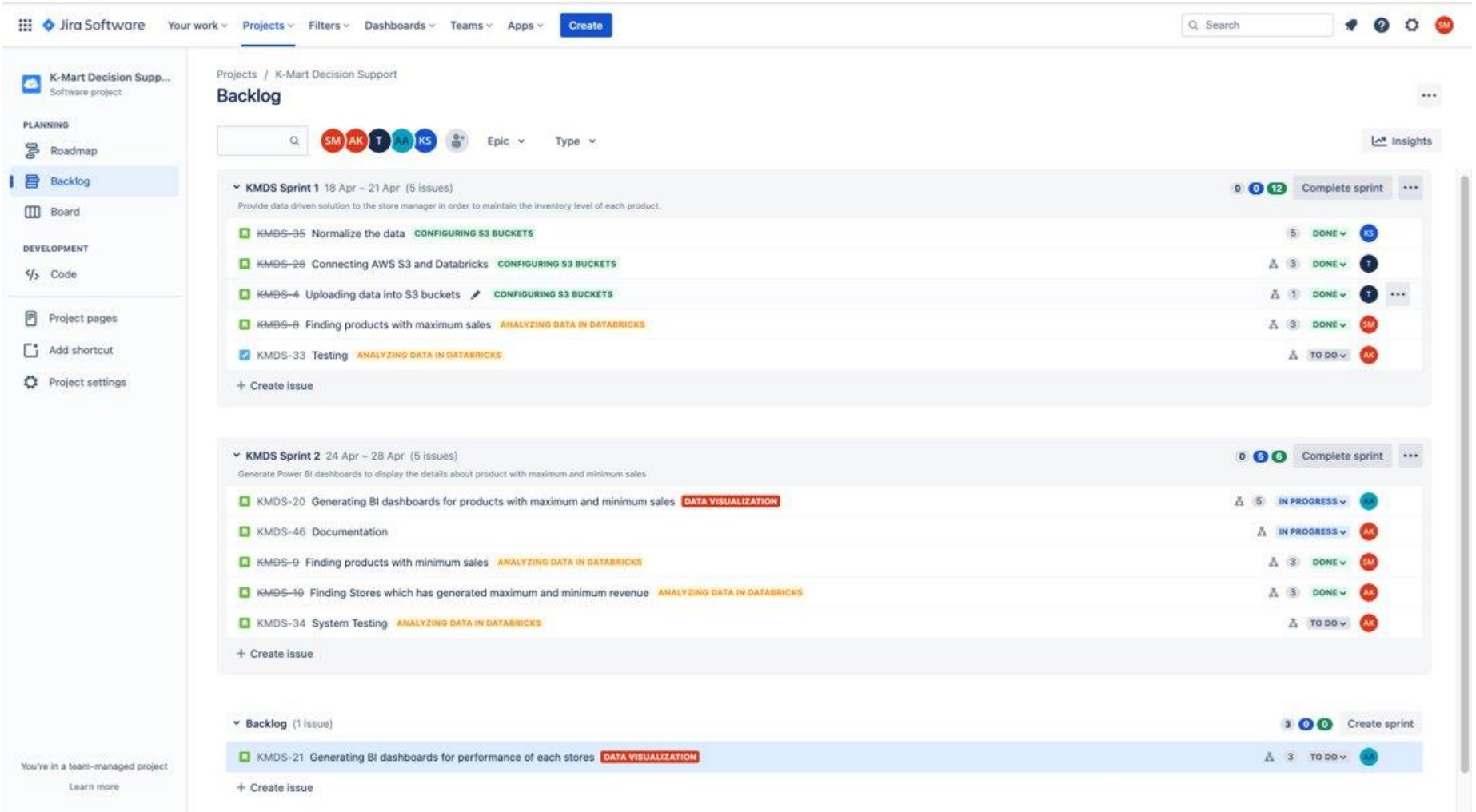
Quering the data In Data Bricks

5

Using Power BI

6

JIRA


Jira Software
Your work
Projects
Filters
Dashboards
Teams
Apps
Create
Search
K-Mart Decision Supp...
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
You're in a team-managed project
Learn more
Projects / K-Mart Decision Support
Backlog
Epic
Type
Insights
KMDS Sprint 1 18 Apr – 21 Apr (5 issues)
Provide data driven solution to the store manager in order to maintain the inventory level of each product.
Complete sprint
KMDS-35 Normalize the data CONFIGURING S3 BUCKETS DONE
KMDS-28 Connecting AWS S3 and Databricks CONFIGURING S3 BUCKETS DONE
KMDS-4 Uploading data into S3 buckets CONFIGURING S3 BUCKETS DONE
KMDS-8 Finding products with maximum sales ANALYZING DATA IN DATABRICKS DONE
KMDS-33 Testing ANALYZING DATA IN DATABRICKS TO DO
Create issue
KMDS Sprint 2 24 Apr – 28 Apr (5 issues)
Generate Power BI dashboards to display the details about product with maximum and minimum sales
Complete sprint
KMDS-20 Generating BI dashboards for products with maximum and minimum sales DATA VISUALIZATION IN PROGRESS
KMDS-46 Documentation IN PROGRESS
KMDS-9 Finding products with minimum sales ANALYZING DATA IN DATABRICKS DONE
KMDS-19 Finding Stores which has generated maximum and minimum revenue ANALYZING DATA IN DATABRICKS DONE
KMDS-34 System Testing ANALYZING DATA IN DATABRICKS TO DO
Create issue
Backlog (1 issue)
Create sprint
KMDS-21 Generating BI dashboards for performance of each stores DATA VISUALIZATION TO DO
Create issue

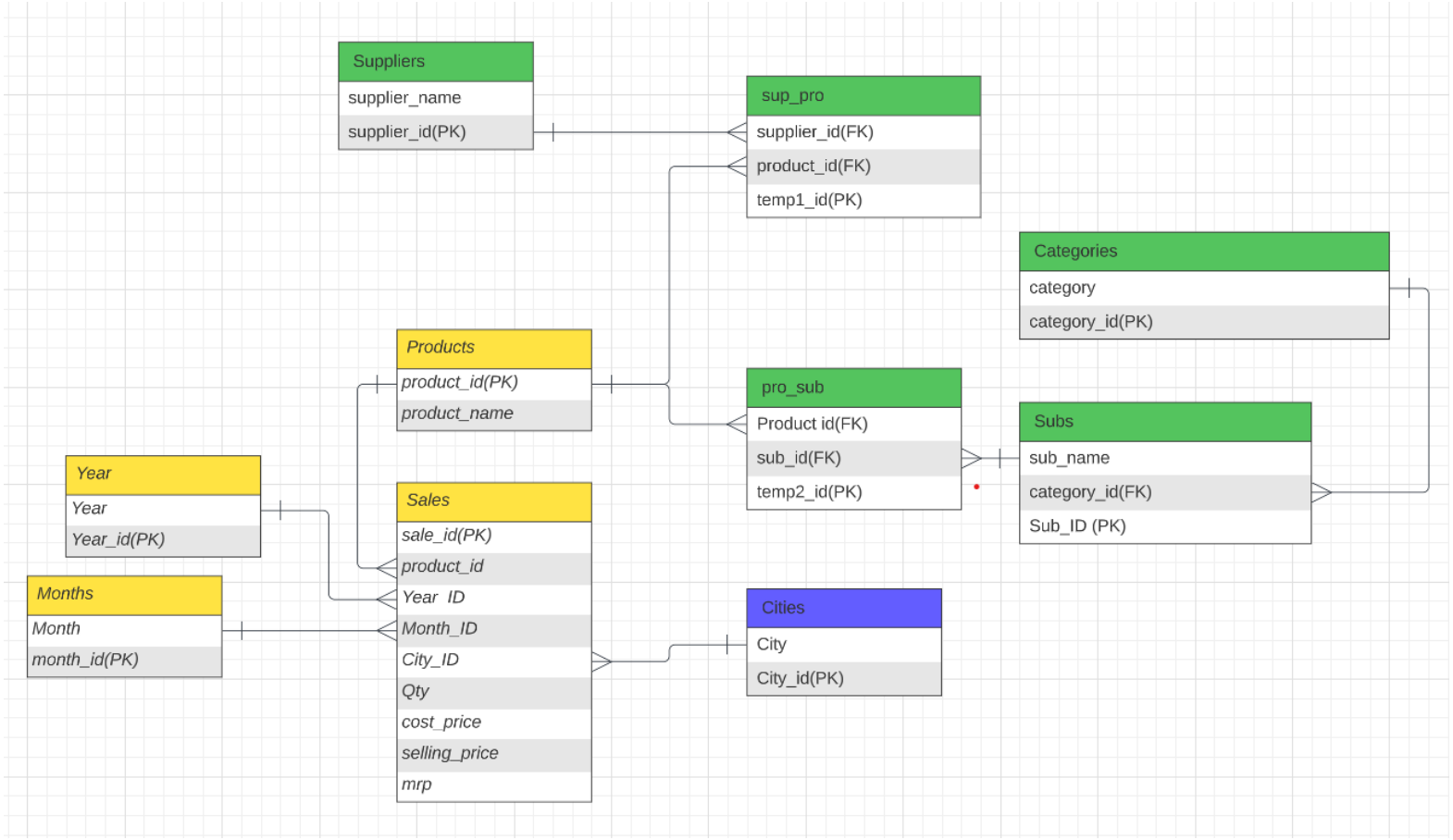
SPRINT 1

User Story	Tech Stack	TASKS	Acceptance Criteria
As a Data Engineer I need to Clean the data .	Data Bricks	• Deleting repeated row's • Removing NULL value from all the column's • Checking for data integrity	• No repeated row's • No NULL values
As a Data Engineer I need to normalize the data so that analyst can query the data effectively	AWS S3	• Create tables and establish relationship between them	• No repeated rows of data • Every tables should contain a primary key • Dimension table should be related to fact table

SPRINT 1 CONTINUATION

User Story	Tech Stack	TASKS	Acceptance Criteria
As a data engineer I need to connect AWS S3 to Data Bricks so that SQL developer can access the data	AWS S3, Data Bricks	• Create S3 bucket • Create Access key and Secret access key for a particular IAM account • Mounting S3 into Data Bricks	• Should test the connection establis hed between S3 and Databricks • Proper access from S3 to Databricks without any impediments
As a data engineer I need to upload data in S3 buckets so that data can be retrieved from S3 to data bricks	AWS S3	• Uploading dataset as an object into s3	• The data stored in the S3 bucket should be readily accessible. • The versioning option must enable to avoid overwriting existing d ata

SCHEMA FOR NORMALIZATION



TABLES

SALES

sales_id	Year_ID	Month_ID	Product_ID	City_ID	QTY	COST_PRICE	SP	MRP
1	1	3	1	1	77	1060	1325	1457.5
2	1	3	1	2	87	1060	1325	1457.5
3	1	3	1	3	82	1060	1325	1457.5
4	1	3	1	4	1	1060	1325	1457.5
5	1	3	1	5	2	1060	1325	1457.5
6	1	3	1	6	3	1060	1325	1457.5
7	1	3	1	7	44	1060	1325	1457.5
8	1	3	1	8	95	1060	1325	1457.5
9	1	3	1	9	27	1060	1325	1457.5
10	1	3	1	10	3	1060	1325	1457.5
11	1	1	2	1	18	240	300	330
12	1	1	2	2	37	240	300	330
13	1	1	2	3	75	240	300	330
14	1	1	2	4	99	240	300	330
15	1	1	2	5	19	240	300	330
16	1	1	2	6	83	240	300	330
17	1	1	2	7	21	240	300	330
18	1	1	2	8	79	240	300	330
19	1	1	2	9	6	240	300	330
20	1	1	2	10	76	240	300	330
21	1	1	3	1	6	180	225	247.5
22	1	1	3	2	21	180	225	247.5
23	1	1	3	3	82	180	225	247.5
24	1	1	3	4	19	180	225	247.5

PRODUCTS

Product_ID	Product_Name
1	ZENA HAND WASH 500ML
2	ZACT SMOKERS TOOTHPASTE 150G
3	ZACT SMOKERS TOOTHPASTE 90G
4	ZACT WHITENING TOOTHPASTE 150G
5	ZACT WHITENING TOOTHPASTE 90G
6	YORK BOTTLE BRUSH (4 COL.) 4108
7	YORK DISH BRUSH (4 COL.) 4102
8	YORK DISH BRUSH (4 COL.) 4105
9	YORK FLY CATCHER (ASST. COL.) Z002
10	YORK SCRUBBING BRUSH MAXI (4 COL.) 4001
11	X-AIR BODY WASH FRESH 355ML
12	X-AIR BODY WASH REG. 355ML
13	WIZZ CONC.FABRIC SOFTENER EXOTIC ORCHID 1LT
14	WIZZ CONC.FABRIC SOFTENER FRESH DAISY 1LT
15	WIZZ IRONING WATER PETAL FRESH 1LT
16	WISDOM DENTAL FLOSS MINT WAXED 100M
17	WISDOM DENTAL STICKS MINT 100'S
18	WISDOM M/WASH FRESH EFFECT COOLMINT 500ML
19	WISDOM M/WASH FRESH EFFECT FRESHMINT 500ML
20	WISDOM DENTURE BOX
21	WISDOM DENTURE BRUSH
22	WISDOM TOOTHBRUSH ANGLED QUEST MEDIUM
23	WISDOM TOOTHBRUSH JUNIOR
24	WISDOM TOOTHBRUSH MEDIUM NYLON 212

SUPPLIERS

Supplier_ID	Supplier_name
1	QUALITY CONCEPT LTD
2	MEDIPHARM INTERNATIONAL LTD
3	KILIMANJARO CARVINGS EXPORTS LTD
4	KILIMANJARO CARVING & EXPORTS
5	MEEXUM ENTERPRISES LIMITED
6	MERIT AFRICA LIMITED
7	DEBENHAM & FEAR LIMITED
8	KULAL INTERNATIONAL LTD
9	BRAND IMPORTS LIMITED.
10	RAPRA LIMITED
11	DIVERSEY EASTERN AND CENTRAL A
12	UNILEVER KENYA LIMITED
13	DIVERSEY EASTERN AND CENTRAL AFR
14	KENROID LIMITED
15	FINSBURY TRADING LTD
16	CHANDARIA INDUSTRIES LTD.
17	TRADE ROOTS LIMITED.
18	PZ CUSSONS (E.A) LTD.
19	TROPIKAL BRANDS (AFRIKA) LTD.
20	MISS BEAUTY CO LTD
21	MISS BEAUTY CO. LTD
22	SALA TRADERS
23	ZAWADI TRADINGS LTD.
24	ZAWADI TRADINGS LTD.

TABLES

SUBS

Sub_ID	Sub_name	Category_ID
1	HANDWASH/GEL	1
2	TOOTHPASTE	2
3	DISH WASHERS	3
4	FABRIC SOFTENERS	3
5	DENTAL FLOSS	2
6	MOUTHWASH & FR	2
7	TOOTHBRUSH	2
8	TABLETS/LIQUIDS S	1
9	SHOWER & BATH A	1
10	SURFACE CLEANER	3

CATEGORIES

Category_id	Category
1	PERSONAL CARE
2	ORAL CARE
3	HOME CARE

CITIES

City_ID	City_name
1	PUNE
2	MUMBAI
3	CHENNAI
4	BANGALORE
5	KOCHI
6	DELHI
7	NAGPUR
8	NASHIK
9	VISHAKAPATNAM
10	BHOPAL

MONTHS

Months_II	Months
1	JAN
2	FEB
3	MAR
4	APR
5	MAY
6	JUN
7	JUL
8	AUG
9	SEP
10	OCT
11	NOV
12	DEC

TABLES

SUP_PRO

A	B	C
Temp1_ID	Product_ID	S_ID
1	1	1
2	2	2
3	3	2
4	4	2
5	5	2
6	6	3
7	7	3
8	8	3
9	9	3
10	10	3
11	6	4
12	7	4
13	8	4
14	9	4
15	10	4
16	11	5
17	12	5
18	13	6
19	14	6
20	15	6
21	16	7
22	17	7
23	18	8
24	19	8

YEARS

Year_ID	Year
1	2014
2	2015

PRO_SUB

Temp_ID	Product_I	Sub_ID
1	1	1
2	2	2
3	3	2
4	4	2
5	5	2
6	6	3
7	7	3
8	8	3
9	9	3
10	10	3
11	11	1
12	12	1
13	13	4
14	14	4
15	15	4
16	16	5
17	17	5
18	18	6
19	19	6
20	20	7
21	21	7
22	22	7
23	23	7
24	24	7

CONNECTING S3 TO DATA BRICKS

Cmd 5

```
1  access_key = df.head(1)[0][0]
2  secret_key = df.head(1)[0][1]
```

▸ (2) Spark Jobs

Command took 1.12 seconds -- by krishnasunderofficial@gmail.com at 4/25/2023, 11:12:33 AM on 25april

Cmd 6

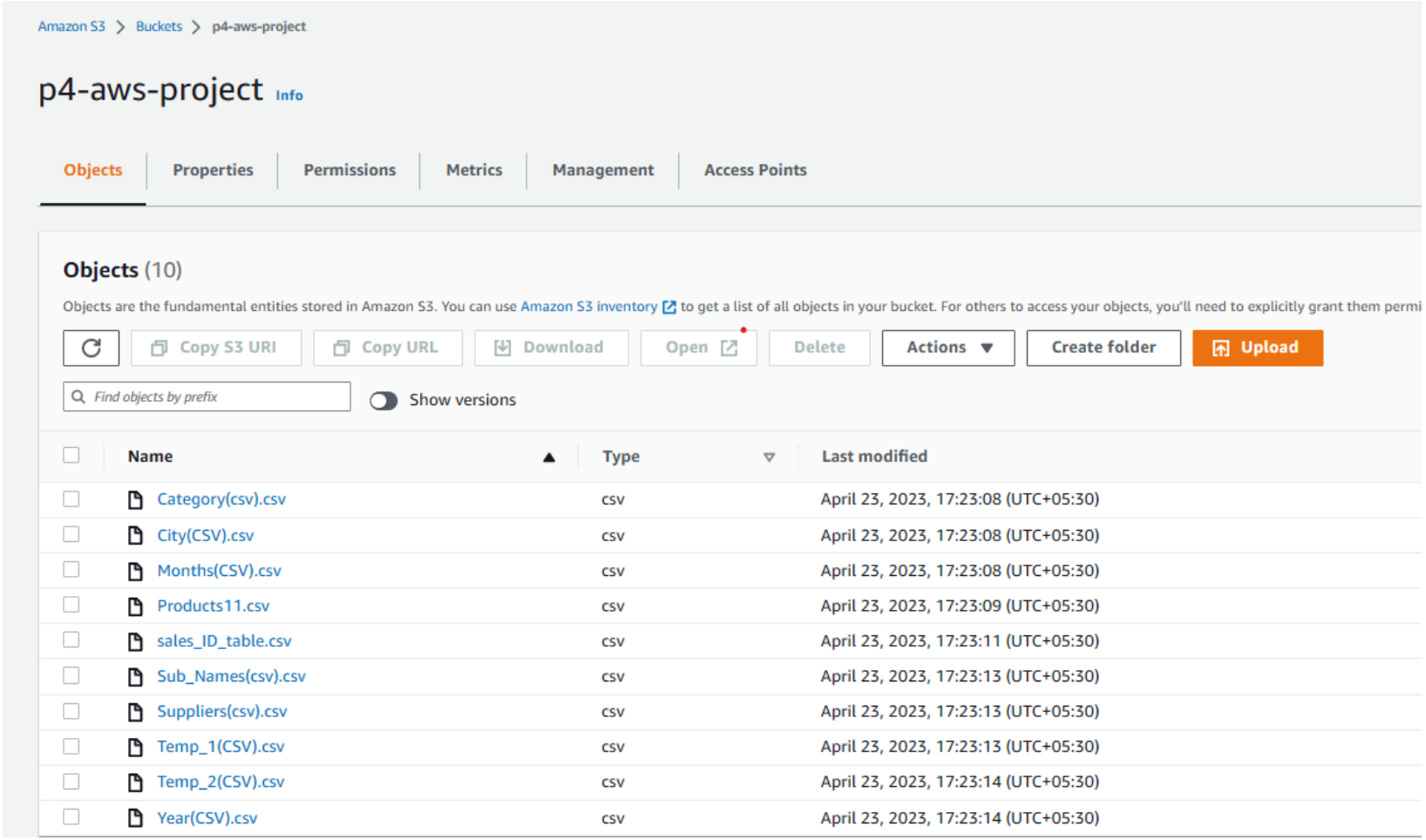
```
1  encoded_secret_key = urllib.parse.quote(secret_key,"")
2  aws_bucket_name = "p4-aws-project"
3  mount_name = "/mnt/s3_p4"
4  source_url = "s3a://{0}:{1}@{2}".format(access_key,encoded_secret_key,aws_bucket_name)
5
6  # Accessing the acess key and the secret key
7  # Creating a source url
```

Command took 0.09 seconds -- by krishnasunderofficial@gmail.com at 4/25/2023, 11:12:33 AM on 25april

Cmd 7

```
1  dbutils.fs.mount(source_url,mount_name)
2
3  # Mounting S3 bucket to Databricks if True connection is established
```

UPLOADING TO S3 BUCKET



THANKYOU

APPENDIX

PRODUCT WITH MAX AND MIN SALE IN A PARTICULAR CITY, YEAR, MONTH AND SUBCATEGORY

	Year_ID	Month_ID	City_ID	Sub_name	Product_Name	sum(QTY)
1	1	1	1	DENTAL FLOSS	ACTIVE D/FLOSS ADVANCED TWIN TRAVEL 12M	86
2	1	1	1	DISH WASHERS	NAKUMATT B/LABEL D/WASHING PASTE LEMON 500G	100
3	1	1	1	FABRIC SOFTENERS	JEYES SO SOFT FAB.COND.TENDER 300ML	100
4	1	1	1	HANDWASH/GEL	ASTONISH HANDWASH WATERLILY 500ML	100
5	1	1	1	MOUTHWASH & FRESHENER	COLGATE PLAX MOUTHWASH WHITENING (BRZ) 500ML	99
6	1	1	1	SHOWER & BATH AGENTS	LYNX SHOWER GEL APOLLO (EU) 250ML	100
7	1	1	1	TABLETS/LIQUIDS SOAP	FLAMINGO SOAP PEACH EXTRACT 3X90G	100

	Year_ID	Month_ID	City_ID	Sub_name	Product_Name	TOTAL_QTY
1	1	1	1	DENTAL FLOSS	WISDOM DENTAL STICKS MINT 100'S	15
2	1	1	1	DISH WASHERS	M/FRESH D/WASH.PASTE ZESTY LEMON 250G	1
3	1	1	1	FABRIC SOFTENERS	C/FRESH TRIGGER FABRIC FRESHENER 500ML	5
4	1	1	1	HANDWASH/GEL	J&J DAILY ESSENT. REFRESHING GEL WASH 150ML	0
5	1	1	1	MOUTHWASH & FRESHENER	LISTERINE MOUTHWASH ZERO (UK) 500ML	1
6	1	1	1	SHOWER & BATH AGENTS	AXE SHOWER GEL HOT FEVER (EU) 250ML	0
7	1	1	1	TABLETS/LIQUIDS SOAP	FA BAR SOAP ALOE VERA 125G	0

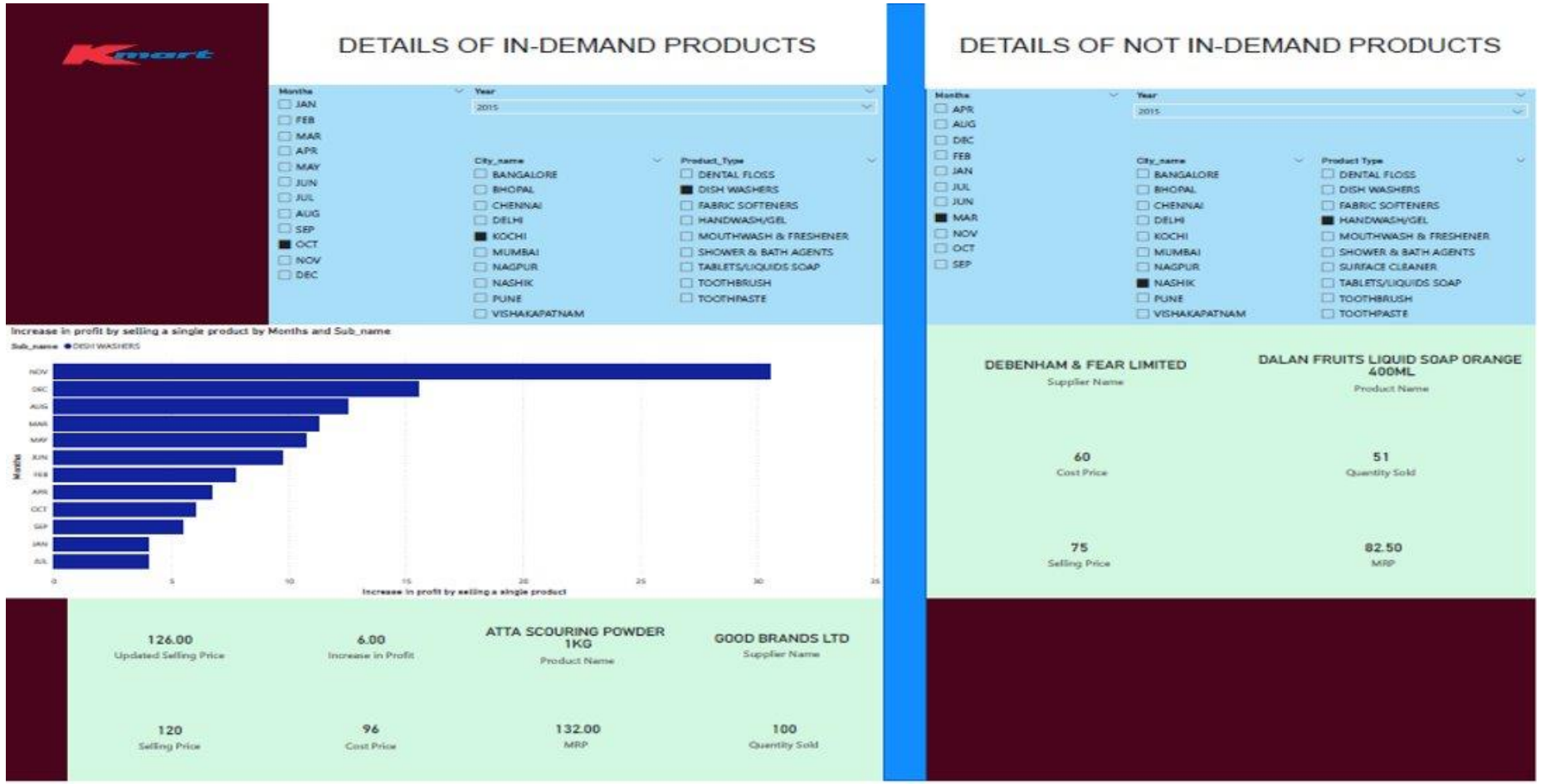
CITIES WITH MINIMUM AND MAXIMUM REVENUE

City_name	revenue
CHENNAI	4.8806587241999984E8
NAGPUR	4.883345966899998E8
PUNE	4.8914286442000014E8
DELHI	4.9116193103999996E8
BANGALORE	4.915273408600001E8
KOCHI	4.916656319000001E8
NASHIK	4.9229743160999995E8
BHOPAL	4.9523293128E8
MUMBAI	4.981255075600003E8
VISHAKAPATNAM	4.9894824901999986E8

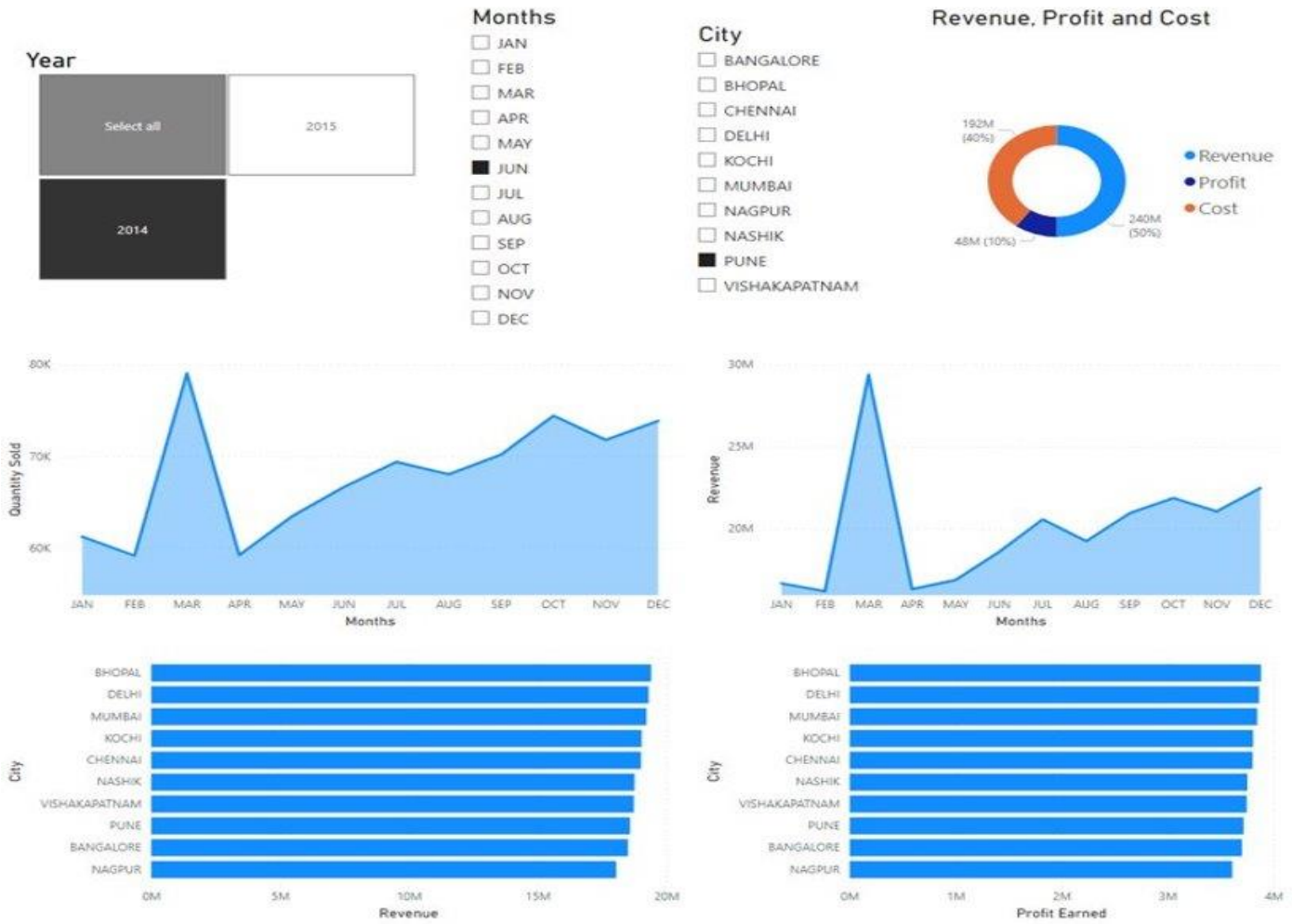
city CHENNAI has the minimum revenue, the revenue being 488065872.41999984

city VISHAKAPATNAM has the maximum revenue, the revenue being 498948249.01999986

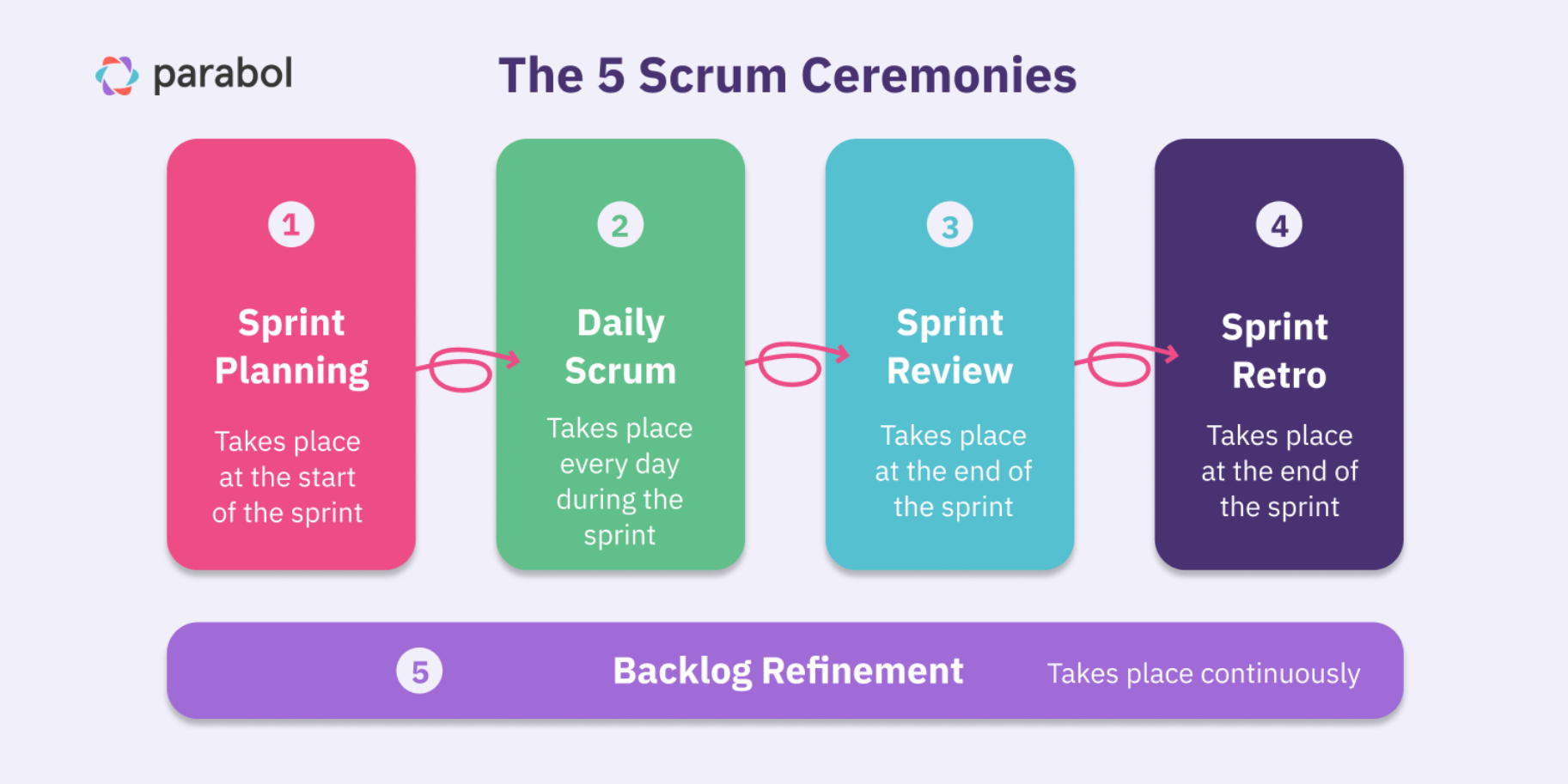
POWER BI DASHBOARD



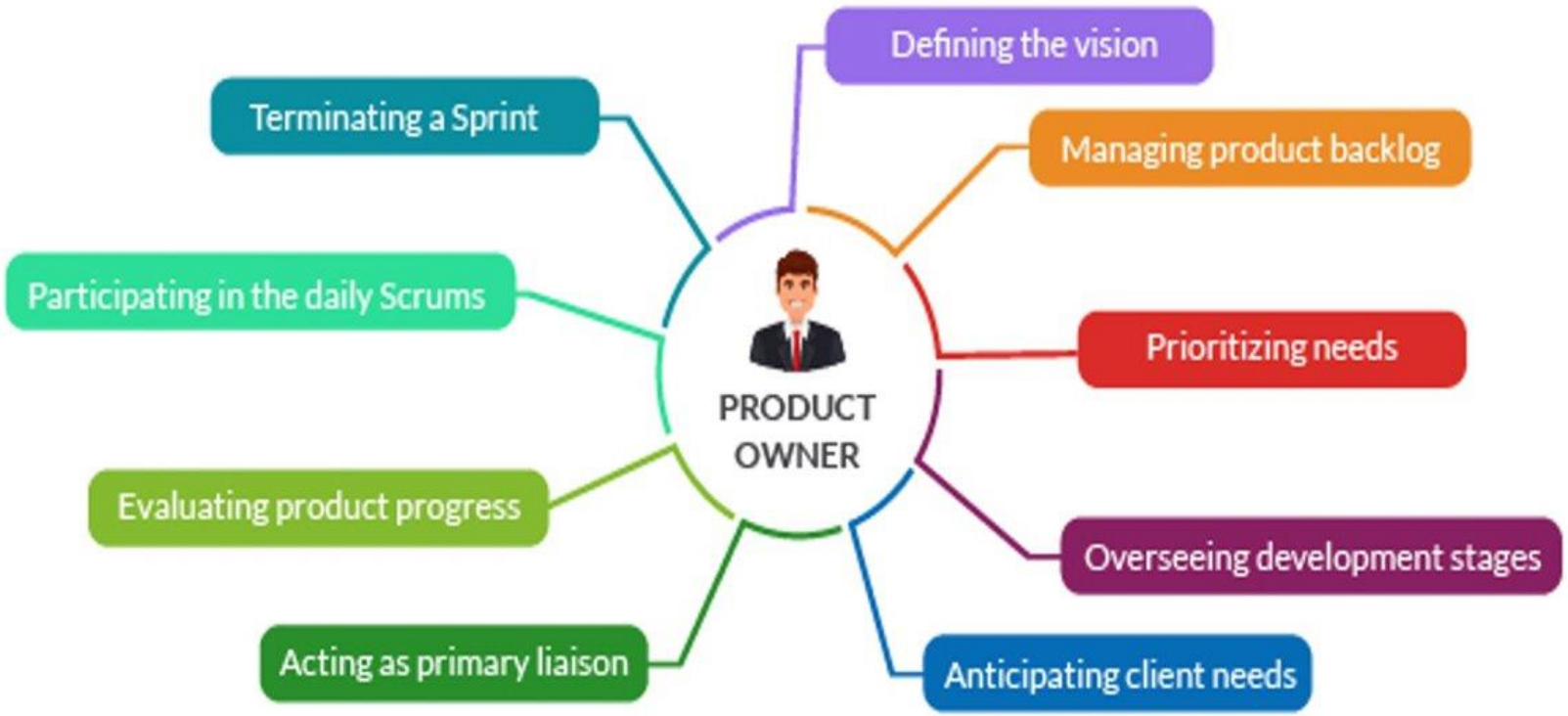
POWER BI DASHBOARD



SCRUM MASTER CEREMONIES



PRODUCT OWNER CEREMONIES



PRODUCT BACKLOG GROOMING

- Removed outdated user stories and tasks.
- Added new user stories that reflect newly discovered user insights.
- Broke down broad user stories into smaller items.
- Reordered user stories based on their priority

- Explained and clearly defined user stories and tasks to avoid uncertainty
- Assigning or re-assigning story points and estimates.
- Identifying roadblocks and minimizing risks related to backlog items.

BEST CODING PRACTICES FOLLOWED

- Naming Convention:
    - Had self-explanatory names for variables and tables.
    - Our function names followed snake casing.
    - Class name followed Camel case.
- Kept Clear and Concise comments.
- Code Indentation was taken care of.
- Followed Reusability and Scalability

"Data will talk to you if you're willing to listen."

- Jim Bergeson

THANK YOU AGAIN