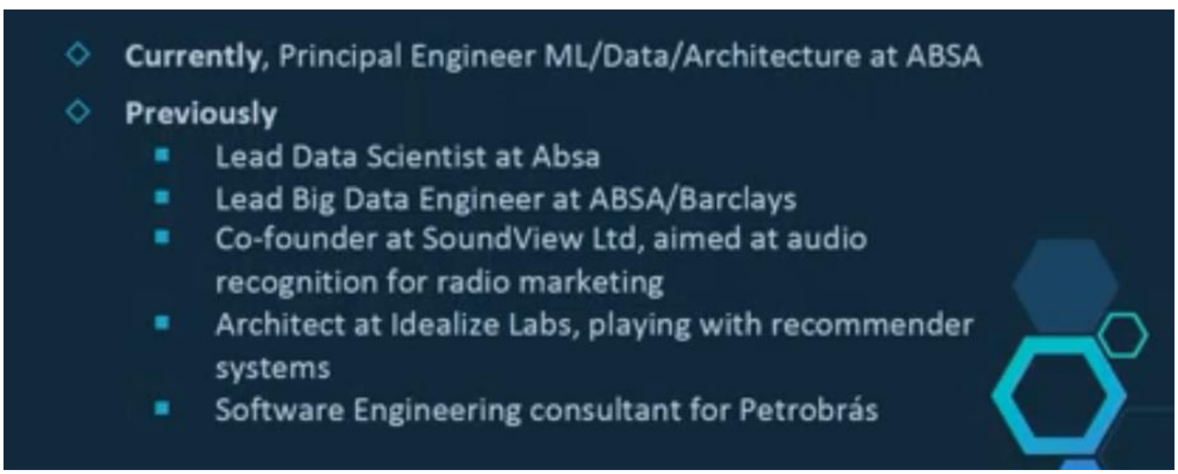
This talk presents the journey we have started at Absa, to transform it into a data-driven organization. It will show the work we've done on the Data Engineering side, migrating from legacy systems into a state-of-the-art technology stack by leveraging and contributing to the open-source space; our transition to a cloud environment; the creation and scaling of a Data Solutions team and how this team has delivered intelligence to different business units; and finally, how all of that have been achieved through architectural changes aimed at creating an MLOps platform horizontal to different analytics and modelling requirements. It will also discuss some of the models we've been building at Absa, from tackling internal automation to improving the relationship with our customer.



Agenda

1. The opportunities
2. The challenges
3. The approaches
4. The new challenges
5. The future

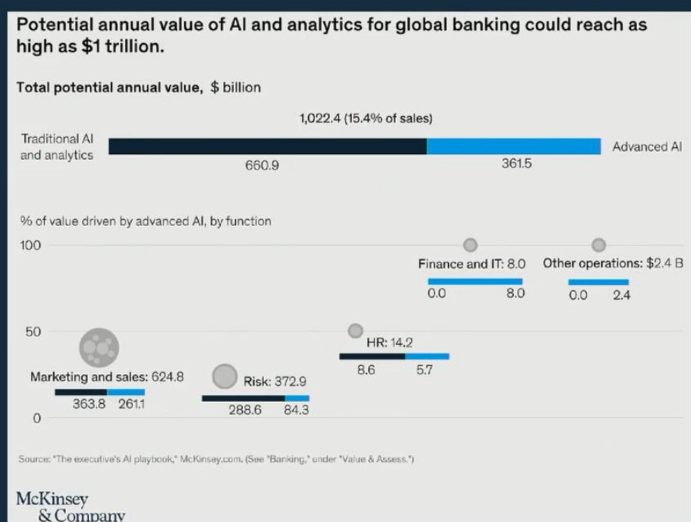
1

The Opportunities

Financials, Positioning and Value Generation

1

The Opportunities - Financial



AI-bank of the future: Can banks meet the AI challenge? https://www.mckinsey.com/industries/financial-services/our-insights/ai-bank-of-the-future-can-banks-meet-the-ai-challenge(2021)

1

The Opportunities - Position

- Decades of data gathering
- Unprecedented increase in data inflow
- Substantial decrease of cost in computing
  - (storage + processing)
- Advances in computational techniques and tools
- Dozens of millions of customers
- Tech-savvy generations

The Opportunities - Value Generation

Big Data

=

Personalization

=

Better product fit

=

Happier customer

=

Loyalty

=

New + Sustained Revenue

8

The Opportunities - Automation

Standardized Processes + AI

=

Automation

=

Better Performance

=

Cost Savings

2 The Challenges

Data Access and Quality, Legacy Systems

2

The Challenges

- Data access and quality
  - Can I access the data?
    - Mainframes
    - Data silos
    - Poor data discoverability
    - Lack of unified analytics / source of truth
  - Can I trust the data?
    - Incompleteness
    - Hacked schemas
    - Duplications
    - Unstructured data
    - Garbage In / Garbage Out
- Workforce
  - Where can talent be found?
  - How can it be scaled?
  - Better buy since can not build?

2

The Challenges – Legacy

- The market for Mainframes is strong, with no signs of cooling down. Mainframes
  - Are used by ***71%*** of ***Fortune 500***
  - Are responsible for ***87%*** of all ***credit card transactions*** in the world
  - Are part of the IT infrastructure of ***92*** out of the ***100 biggest banks*** in the world
  - Handle ***68%*** of the world's production ***IT workloads***, while accounting for only ***6%*** of ***IT costs***.
- For companies relying on Mainframes, becoming data-centric can be prohibitively expensive
  - High cost of hardware
  - Expensive business model for data science related activities

3

The Approaches

Technical, Data Engineering, Science, Feature Generation, Models

3

The Approaches

- **Streamlined Data Engineering**
  - Leveraging open-source solutions
  - Developing tools in-house to cover gaps
- **Cross-domain feature generation**
  - Transactional
  - Savings
  - Loans
    - Can a customer's transactional behavior predict the **need** for a loan?
- **Deep Learning for automation tasks**
- **Streamlined MLOps**
  - Standardize processes for feature generation, modelling and deployment

3.1 Platform

DataOps, MLOps, open-source, etc

3.1

Ops Challenges

- We have different use cases ...
  - Batch
  - Online
- ... and different functions
  - Data Engineering
  - Data Science
  - Data Analysis
  - DevOps
- ... but we need a centralized platform that ...
  - Connects the functions seamlessly
  - Serves all use cases
  - Clarifies the trade-off Buy vs Build
    - **SPOILER: introduce a Tooling team**

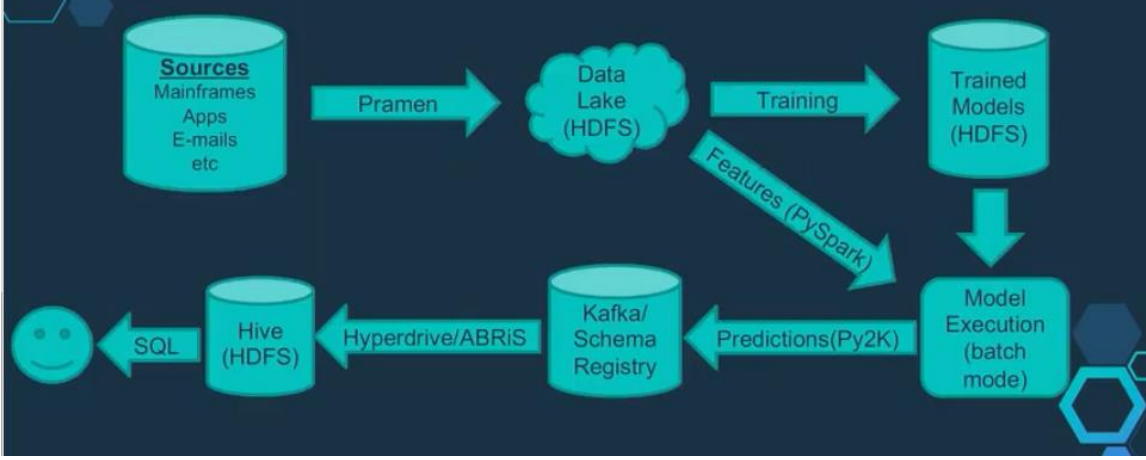
3.1

In-house tooling, first batch

- **Pramen**
  - Spark framework to support data ingestion and quality
- **Py2K**
  - Python library to connect pandas, Kafka and Schema Registry
- **ABRiS**
  - Scala library to connect Spark, Kafka and Schema Registry
- **Hyperdrive**
  - Spark framework for streaming data manipulation
- **Subatomic**
  - Automates the integration between repos and CI/CD tools
- **FeatureMaker**
  - PySpark framework to generate features
- **ModelMaker**
  - Python framework to support model building

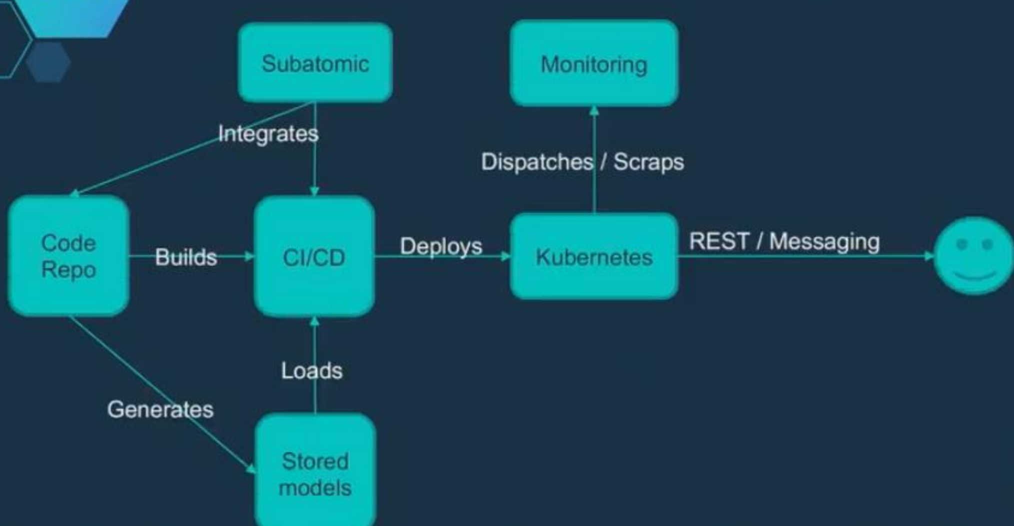
3.1

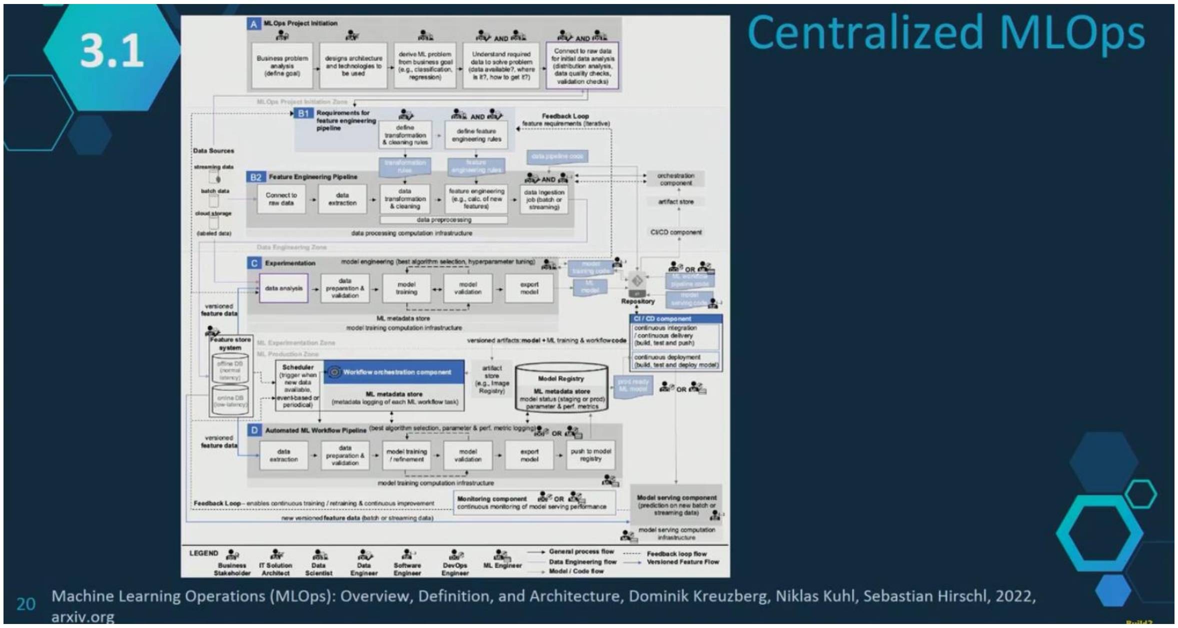
Data/MLOps on-prem - Batch

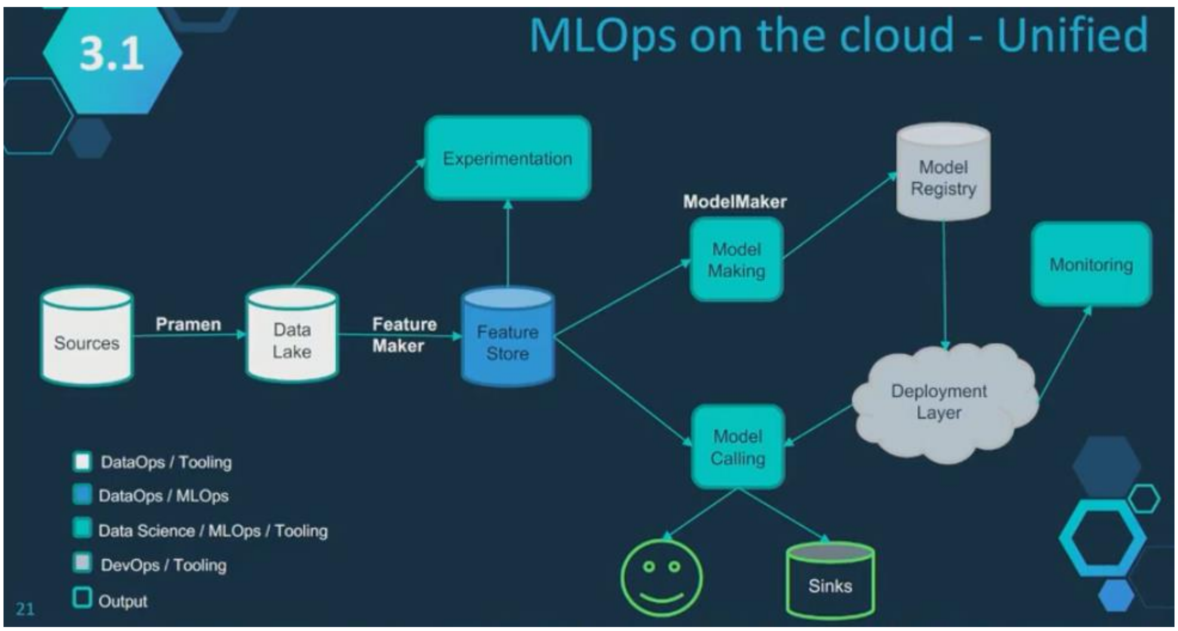


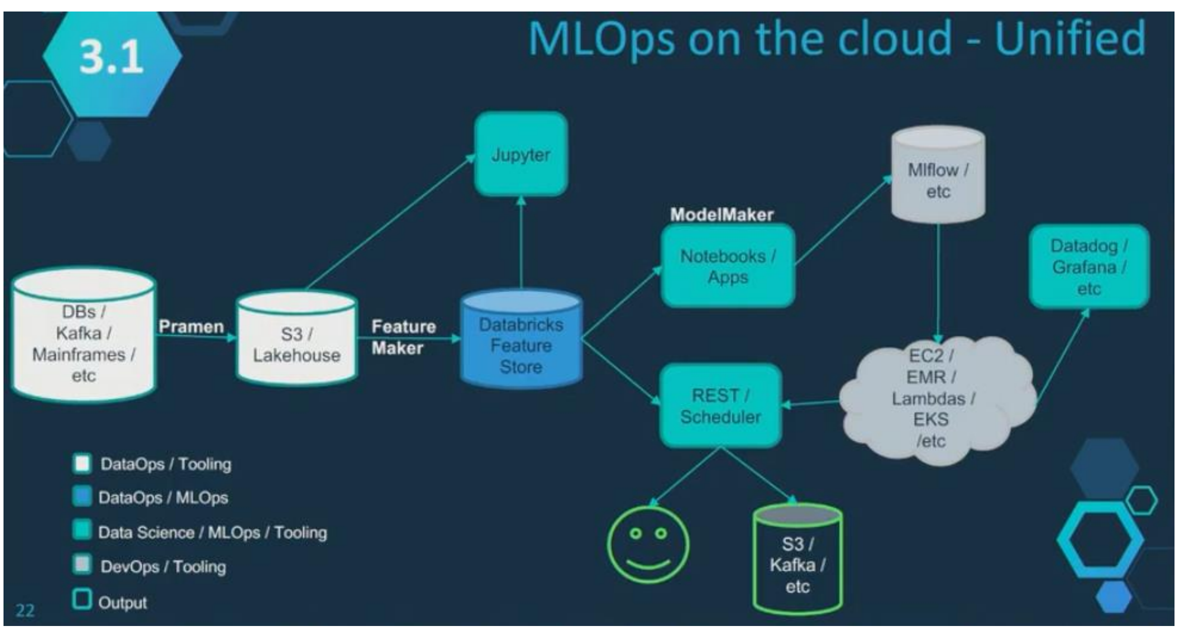
3.1

Data/MLOps on-prem - Online




3.1
Centralized MLOps
20
Machine Learning Operations (MLOps): Overview, Definition, and Architecture, Dominik Kreuzberg, Niklas Kuhl, Sebastian Hirschl, 2022, arxiv.org


3.1
MLOps on the cloud - Unified
Sources
Pramen
Data Lake
Feature Maker
Feature Store
Experimentation
ModelMaker
Model Making
Model Registry
Monitoring
Deployment Layer
Model Calling
Sinks
DataOps / Tooling
DataOps / MLOps
Data Science / MLOps / Tooling
DevOps / Tooling
Output
21


3.1
MLOps on the cloud - Unified
DBs / Kafka / Mainframes / etc
Pramen
S3 / Lakehouse
Feature Maker
Databricks Feature Store
Jupyter
ModelMaker
Notebooks / Apps
Mlflow / etc
Datadog / Grafana / etc
EC2 / EMR / Lambdas / EKS /etc
REST / Scheduler
S3 / Kafka / etc
DataOps / Tooling
DataOps / MLOps
Data Science / MLOps / Tooling
DevOps / Tooling
Output
22

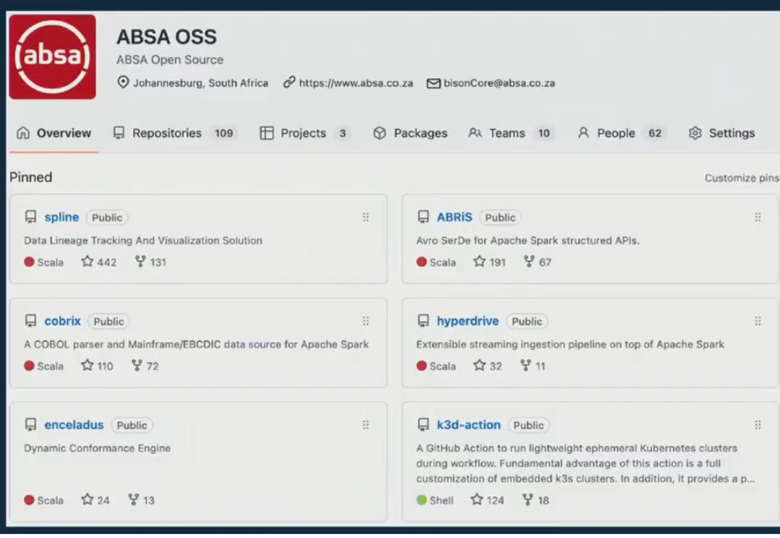
3.1

Data Engineering Challenges

- Available tools not generic enough
  - Cobol processing
- Vendor lock-in
  - Mainframes
- Price
- Gaps in open-source space

3.1

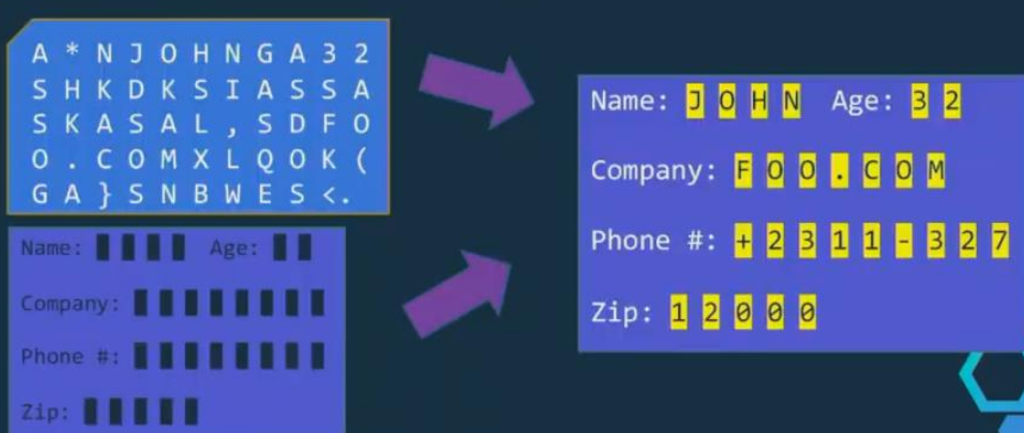
In-house tooling, second batch

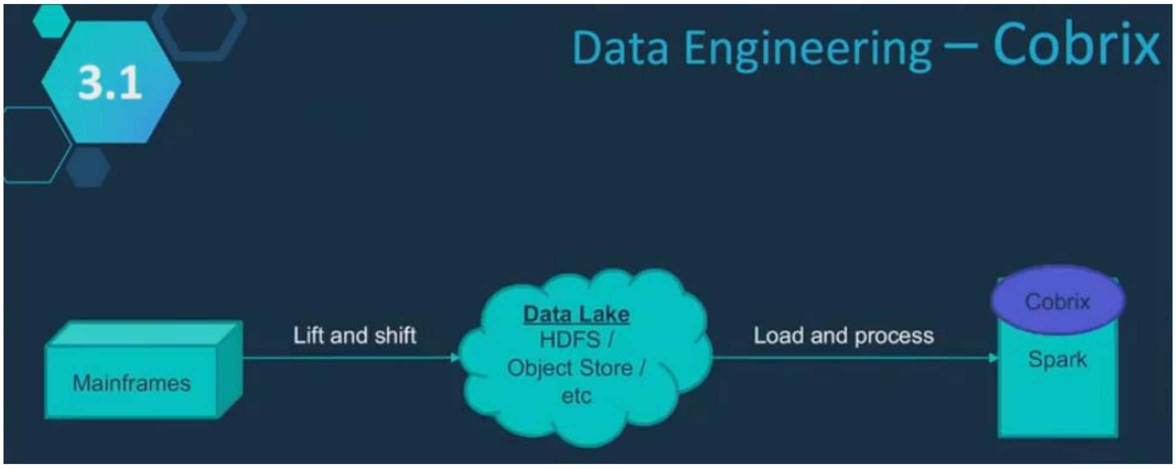


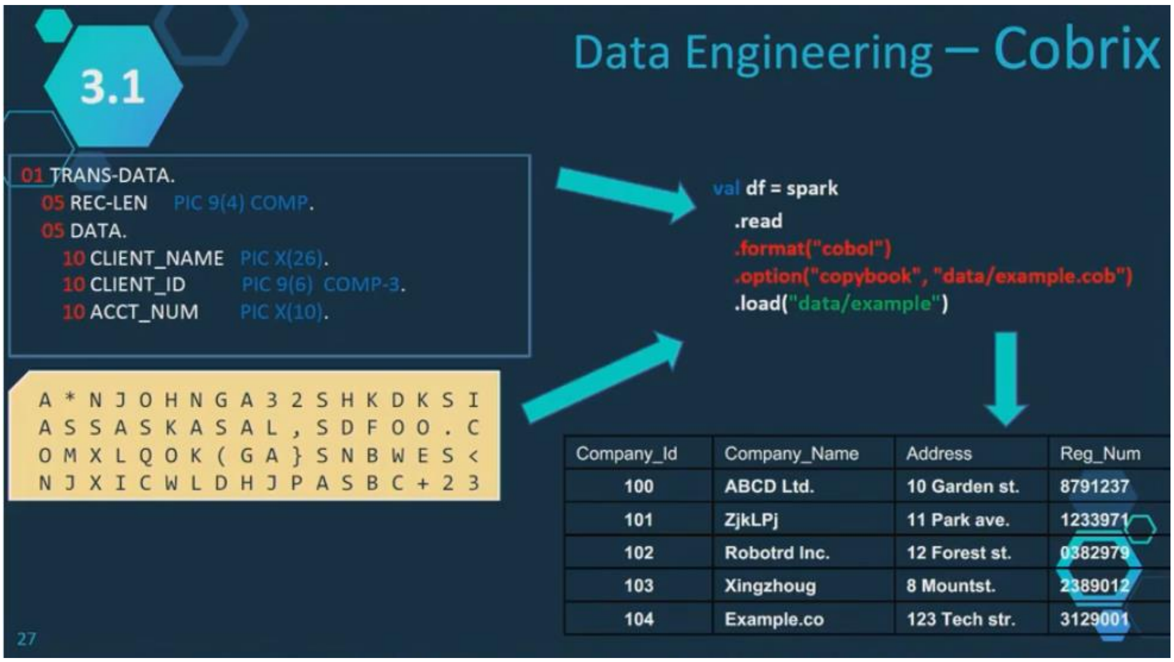
3.1

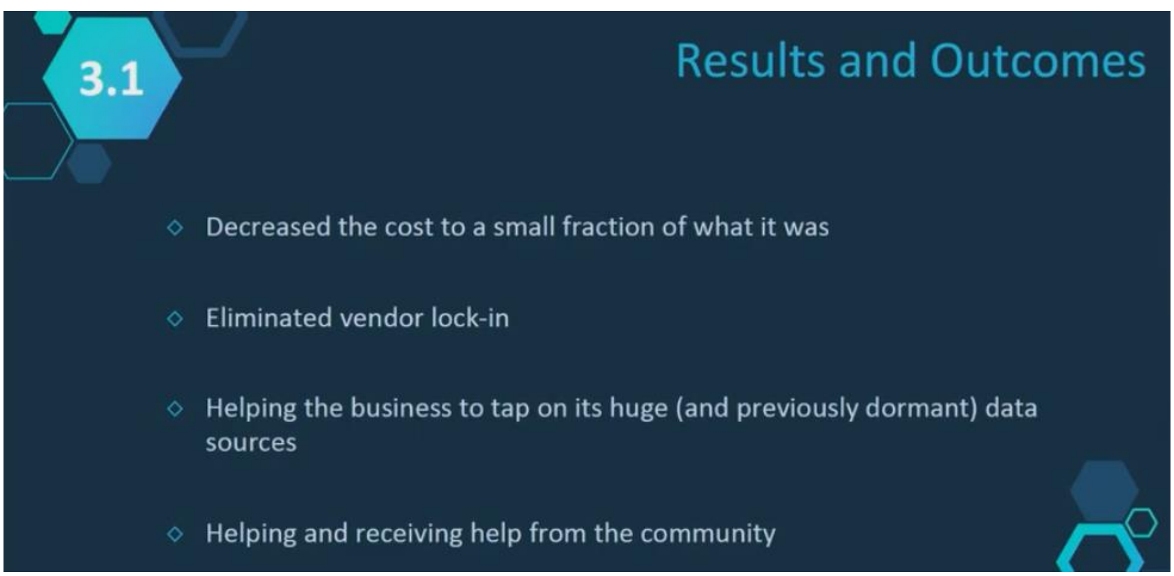
Data Engineering – Cobrix

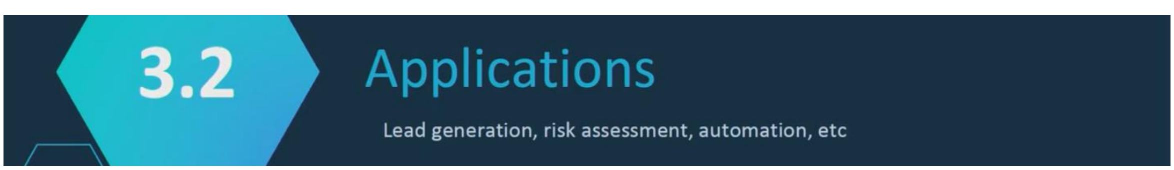
- A **data file** is a collection of records
- A **copybook** is a schema definition




3.1
Data Engineering – Cobrix
Mainframes
Lift and shift
Data Lake
HDFS /
Object Store /
etc
Load and process
Cobrix
Spark

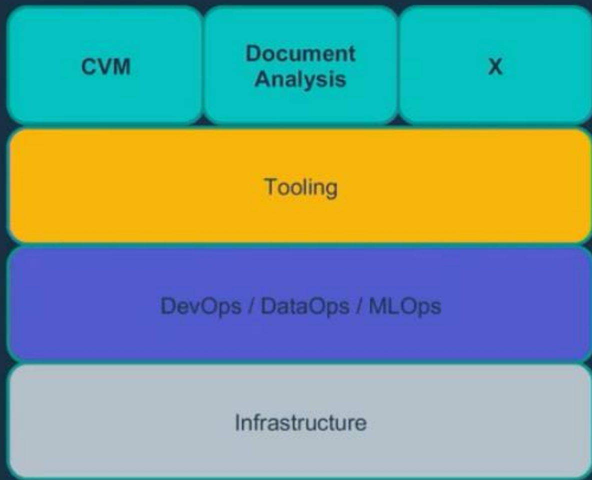

3.1
Data Engineering – Cobrix
01 TRANS-DATA.
05 REC-LEN PIC 9(4) COMP.
05 DATA.
10 CLIENT_NAME PIC X(26).
10 CLIENT_ID PIC 9(6) COMP-3.
10 ACCT_NUM PIC X(10).
A * N J O H N G A 3 2 S H K D K S I
A S S A S K A S A L , S D F O O . C
O M X L Q O K (G A } S N B W E S <
N J X I C W L D H J P A S B C + 2 3
val df = spark
.read
.format("cobol")
.option("copybook", "data/example.cob")
.load("data/example")
Company_Id Company_Name Address Reg_Num
100 ABCD Ltd. 10 Garden st. 8791237
101 ZjkLPj 11 Park ave. 1233971
102 Robotrd Inc. 12 Forest st. 0382979
103 Xingzhoug 8 Mountst. 2389012
104 Example.co 123 Tech str. 3129001
27


3.1
Results and Outcomes
Decreased the cost to a small fraction of what it was
Eliminated vendor lock-in
Helping the business to tap on its huge (and previously dormant) data sources
Helping and receiving help from the community


3.2
Applications
Lead generation, risk assessment, automation, etc

3.2

Data Solutions Space



3.2

CVM – Customer Value Management

- What it does
  - Customer understanding
  - Lead generation
- What are the challenges?
  - Modelling
  - Feature engineering
    - Discoverability
    - Lineage
  - Explainability
  - Time-to-market

3.2

Science

- Gradient Boosted trees (XGBoost, Catboost, etc) + Tabular Features = **SUCCESS**

```
@misc{
    shwartzziv2021tabular,
     title={
       Tabular Data: Deep Learning is Not All You Need
     },
     author={Ravid Shwartz-Ziv and Amitai Armon},
     year={2021},
     eprint={2106.03253},
     archivePrefix={arXiv},
     primaryClass={cs.LG}
}
```

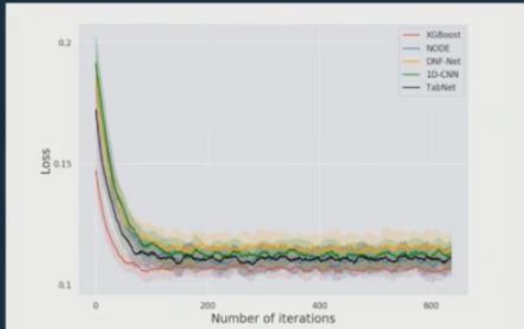


Figure 2: **The Hyper-parameters optimization process for different models.**

3.2

Data-Centric / AutoML

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
gbc	Gradient Boosting Classifier	0.7676	0.7975	0.9128	0.7771	0.8395	0.4286	0.4472
catboost	CatBoost Classifier	0.7645	0.7964	0.9028	0.7787	0.8362	0.4258	0.4408
lightgbm	Light Gradient Boosting Machine	0.7614	0.7888	0.8952	0.7792	0.8332	0.4214	0.4341
ada	Ada Boost Classifier	0.7611	0.7912	0.8968	0.7781	0.8333	0.4195	0.4329
et	Extra Trees Classifier	0.7602	0.7806	0.8984	0.7764	0.8330	0.4158	0.4301
rf	Random Forest Classifier	0.7600	0.7812	0.8990	0.7760	0.8330	0.4151	0.4295
xgboost	Extreme Gradient Boosting	0.7499	0.7713	0.8741	0.7777	0.8231	0.4015	0.4096
lr	Logistic Regression	0.7430	0.7760	0.8583	0.7784	0.8164	0.3919	0.3973
lda	Linear Discriminant Analysis	0.7367	0.7707	0.8556	0.7730	0.8122	0.3755	0.3810
ridge	Ridge Classifier	0.7362	0.0000	0.8659	0.7676	0.8138	0.3674	0.3754
knn	K Neighbors Classifier	0.7175	0.7155	0.8375	0.7617	0.7978	0.3325	0.3368
svm	SVM - Linear Kernel	0.7105	0.0000	0.8342	0.7562	0.7931	0.3147	0.3194
nb	Naive Bayes	0.6797	0.7390	0.6819	0.8105	0.7342	0.3347	0.3481
dt	Decision Tree Classifier	0.6654	0.6294	0.7380	0.7541	0.7459	0.2560	0.2562
qda	Quadratic Discriminant Analysis	0.6247	0.7341	0.7415	0.7445	0.6798	0.1362	0.2068

3.2

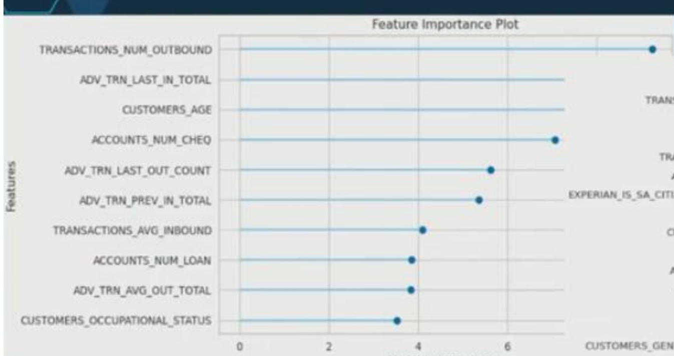
Analytics / Reusable Features

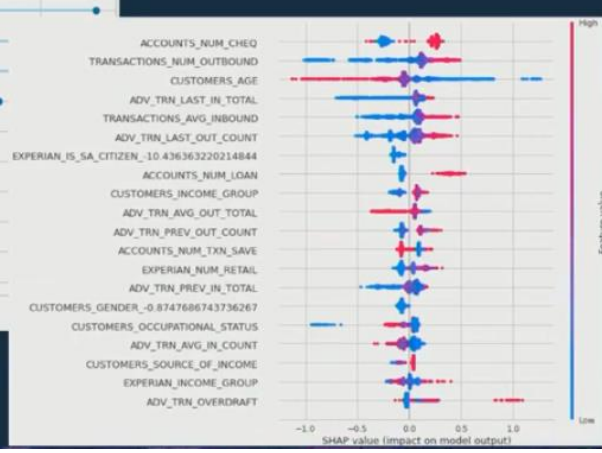
customerKey	
accountNumber	
predictedIncome	46,500.00
avgInfluxTotal	15,366.78
avgOutfluxTotal	14,903.72
previousInfluxTotal	14,500.00
previousOutfluxTotal	15,075.12
lastInfluxTotal	1,125.00
lastOutfluxTotal	18,397.70
percFromAvgInTotal	(0.93)
percFromAvgOutTotal	0.23
avgInterMonthInfluxVar	0.04
avgInterMonthOutfluxVar	0.01
financial_capability	463.06
risk	0.73

- **High degree of change in last period**
  - -93% [% Change of Previous Inflow to Average Inflow Total]
    - Reduction of 93% inflow into account
  - 23% [% Change of Previous Outflow to Average Outflow Total]
    - Increase of 23% outflow into account
- **High Confidence to Interpret**
  - 4% [Average InterMonth Inflow Fluctuations]
    - Stable inflow into account for 6 months
  - 1% [Average InterMonth Outflow Fluctuations]
    - Stable outflow into account for 6 months

3.2

Explainability





3.2 Inter-domain features = faster modelling

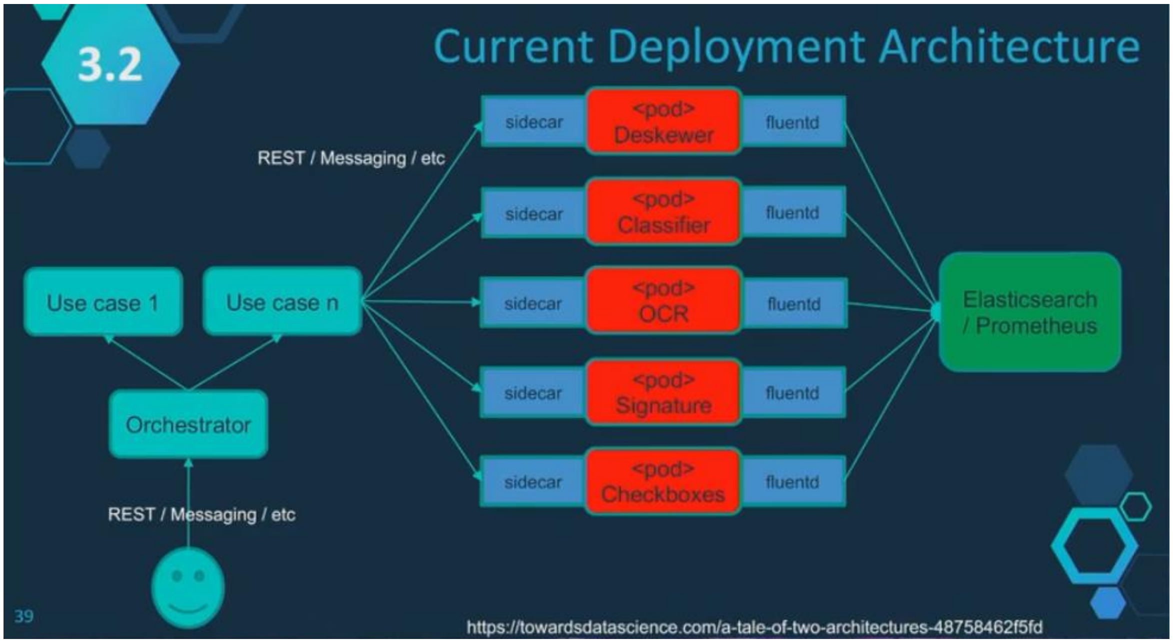
- **Propensity models (2 – 4 weeks lead time)**
  - New to Card
  - New to Transactional
  - New to Overdraft
  - New to Digital
  - New to Personal Loans
  - New to Savings & Investments
  - New to ...

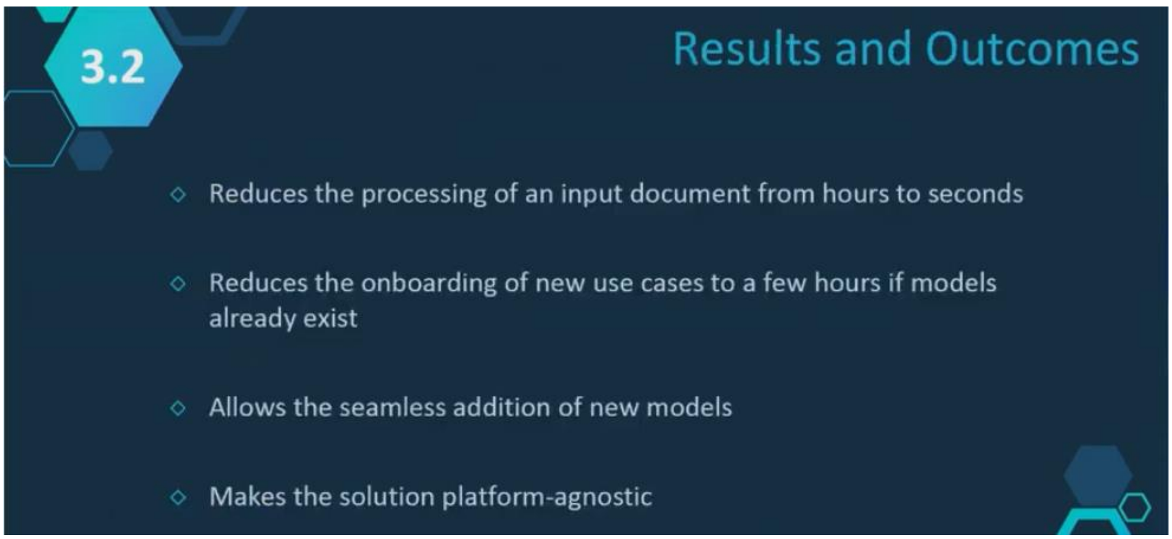
3.2 Results and Outcomes

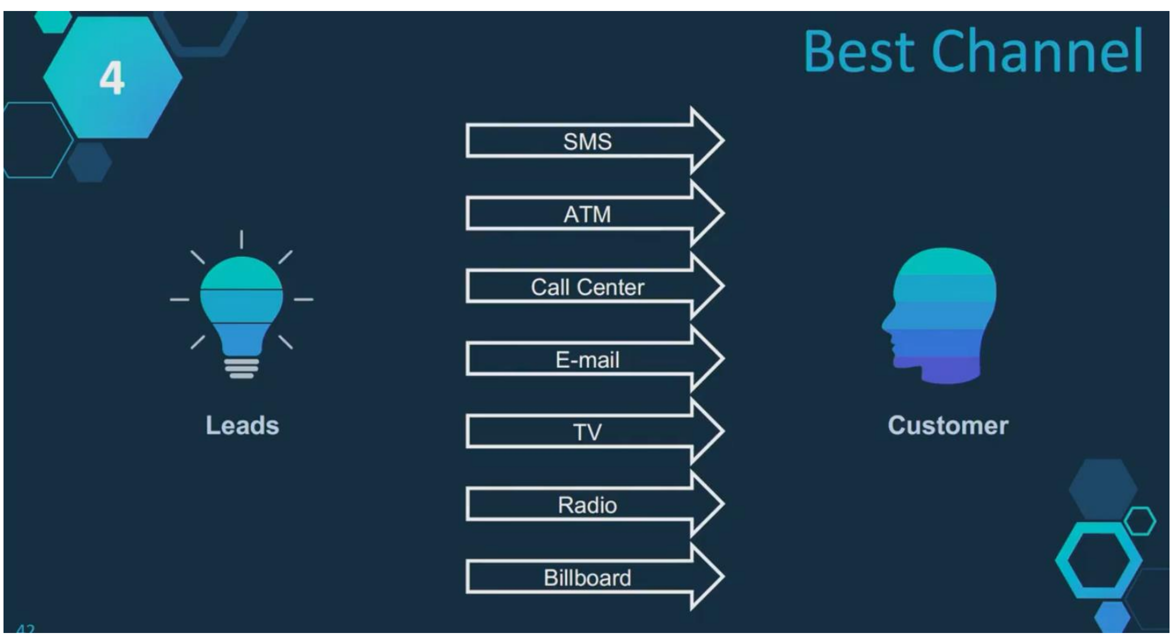
- On average, response increased 300%
- Improves the take-up rate by 2-15% depending on the use case
- Models deployed end-to-end in 1 month or less
- Helping the business to integrate its processes with AI, thus, becoming more data-driven

3.2 Document Analysis Challenges

- **Use case**
  - Business units receive forms that can be filled by hand, scanned, photographed, etc, along with documents like id cards, birth certificates, etc
  - Textual values, checkboxes and signatures need to be identified and extracted
  - Human errors are inevitable
- **Solution**: Automate the extraction
- **Challenges**
  - Skewed images
  - Classification of different documents with similar templates
  - Lack of labelled data
  - SLAs
  - Different types of AI tasks

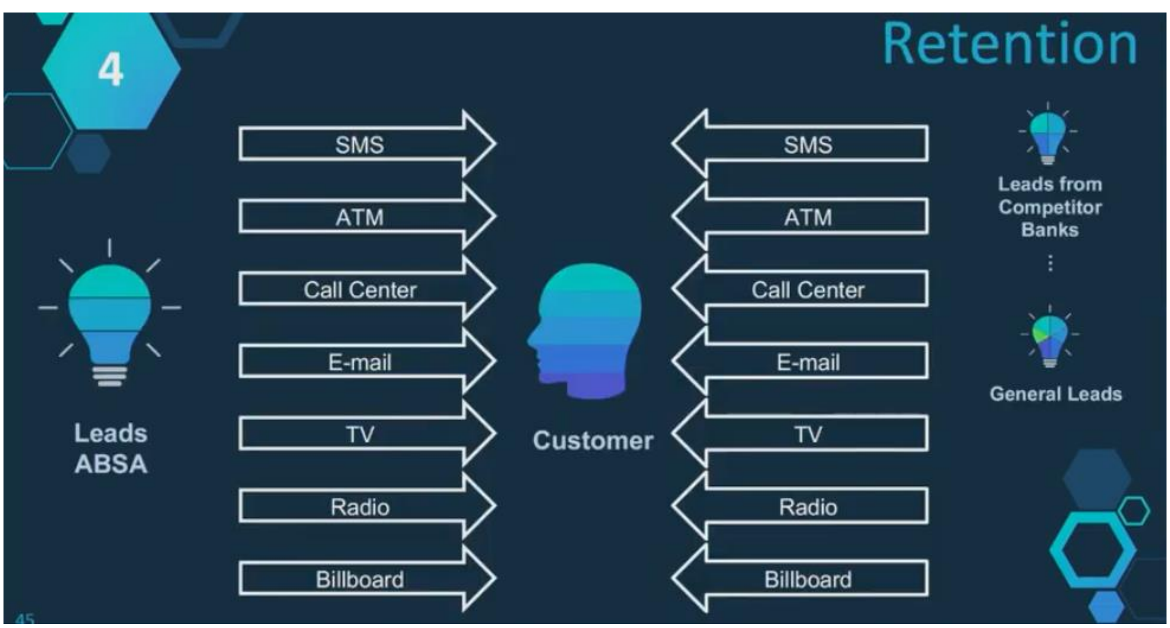

3.2
Current Deployment Architecture
REST / Messaging / etc
sidecar
<pod>
Deskewer
fluentd
sidecar
<pod>
Classifier
fluentd
sidecar
<pod>
OCR
fluentd
sidecar
<pod>
Signature
fluentd
sidecar
<pod>
Checkboxes
fluentd
Elasticsearch / Prometheus
Use case 1
Use case n
Orchestrator
REST / Messaging / etc
39
https://towardsdatascience.com/a-tale-of-two-architectures-48758462f5fd


3.2
Results and Outcomes
◇ Reduces the processing of an input document from hours to seconds
◇ Reduces the onboarding of new use cases to a few hours if models already exist
◇ Allows the seamless addition of new models
◇ Makes the solution platform-agnostic


4
Best Channel
SMS
ATM
Call Center
E-mail
TV
Radio
Billboard
Leads
Customer


4
Lead Selection
Increase in number of sales
400
350
300
250
200
150
100
50
0
Propensity = 0.4
0
50000
100000
150000
200000
250000
300000
Number of Leads Executed
Increase in Sales


4
Marketing Consent
Number of TXN Customers with Marketing Consent
Millions
16
14
12
10
8
6
4
2
-
r = -0.77
3.58
3.57
3.56
3.55
3.54
3.53
3.52
3.51
Millions
Number of TXN Leads Executed
2020/01/01
2020/02/01
2020/03/01
2020/04/01
2020/05/01
2020/06/01
2020/07/01
2020/08/01
2020/09/01
2020/10/01
2020/11/01
2020/12/01
2021/01/01
2021/02/01
2021/03/01
2021/04/01
2021/05/01
Number of TXN Leads Executed
Number of Customers with Marketting Consent


4
Retention
Leads ABSA
SMS
ATM
Call Center
E-mail
TV
Radio
Billboard
Customer
SMS
ATM
Call Center
E-mail
TV
Radio
Billboard
Leads from Competitor Banks
⋮
General Leads


5
The Future
Keep Watching, Beyond Banking


5
Keep Watching
Better Offers
Take-up
Models
Customer
Behavior
KEEP LEARNING


5
Beyond Banking
Credit Card
Loan
Café
Holiday Trip
Better Offers
Take-up
Rewards
Discounts
Models
Customer
Behavior
KEEP LEARNING


The best way to
predict the future is
to create it.
Abraham Lincoln

Agenda

1. The opportunities
2. The challenges
3. The approaches
4. The problems
5. The future

Special thanks to:

ABSA CTO, Data Solutions space, Tooling and Cloud Teams ...