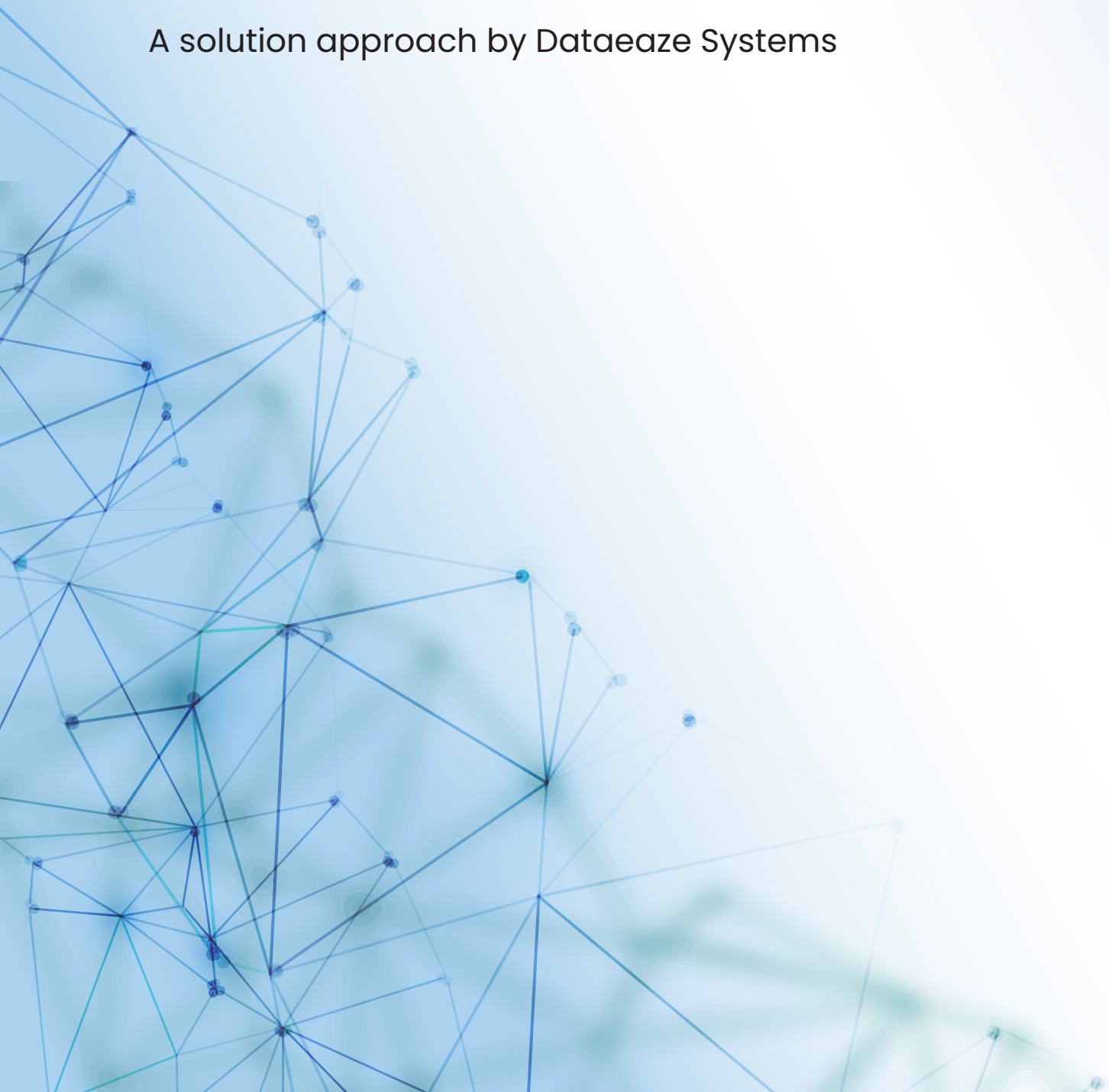
dataeaze

Case Study

# Data platform for large corporate and consumer finance corporation

A solution approach by Dataeaze Systems



# About the client

Client is a leading organization. It is a pan-India credit platform with core businesses in corporate lending, housing finance, digital consumer and MSME finance and asset management.
Client is a lending backend for 20+ other different fintech companies with micro finance focus.

# About Data

Being a company with lending business, client has following functionalities within,

- Loan origination system
- Loan management system
- Collections
- Risk and analytics
- Lead generation

Salesforce is their primary system where multiple of above functionalities are served from, there are other RDBMS and noSQL data sources as well hosting data.

# Need of Analytics

## Need of reliable single source of consolidated data

Data is spread across multiple systems,
- Salesforce as LOS, LMS and lead management
- Mysql, files, third party APIs as other sources of data for other functions

In order to enable any meaningful consolidated analytics, it is required to have data from all sources into a single central data lake and warehouse.

Data onboarded should be accurate and reliable. It is important to have obserability set in order to ensure reliability of data. It is expected that data is to be consolidated into S3 and every other further analytics to refer this for analytics and data reconciliation. Every analytics should not go to source salesforce through APIs / reports. Most of the use cases to be served from this data lake layer

## Near real time refresh of key data elements

For the purpose in order to achieve accurate and latest state of business (specially collections and some other reports from other functions), it was required to have data pulled from salesforce into the data warehouse at a frequency gap of less than 15 minutes.
This is with reliable and accurate data pull, running continuously for 100+ objects and with minimal maintenance (single person, part time should be able to support continuous executions).

### Extendability of pipelines : Low cost and configurable

LMS and LOS tool (salesforce) allows considerable customization, which give agility to client's business. This has an impact that there are frequent improvements in salesforce object schemas (fields getting added / updated).
It is expected from a data platform,

- It should be completely configurable in case of 'adding new object, 'update to existing data getting onboarded', 'addition of columns' etc.
- It should be capable of pulling and processing 100+ objects smoothly with a gap of < 15 minutes. Which should happen in an efficient way with minimal maintenance efforts.

### Various internal analytics reports

Multiple department use data warehouse on redshift as well as Athena in order to achieve analytics dashboards (tableau) and reports. It enables collection, accounts, legal, risk and analytics in order to build their own dashboards and reports.

### Customer data platform

Single central data mart data model to host customer specific data at single location. This is in roadmap.
Achieve use cases like : sales funnel analysis, write off reports etc. get enabled with a single central data mart with customer and customer transaction information.

### Data warehouse with slowly changing dimensions

Customer had a requirement to achieve data warehouse with SCD, so that any updates on previous data are easy to track. Same was implemented within redshift.

## Customer pain points

### Near real time data pull from salesforce is not trivial

- Need to ensure no upper limit reach for API consumption
- There are 100+ objects, achieve robustness of all jobs running at 5min to 15min internval

### History data pull since beginning,

- Since data being from salesforce, historical analytics becomes very salesforce reporting interface specific
- Solution is to pull data to single central source of truth data lake and data warehouse.

### Robust and incremental update to data warehouse with slowly changing dimensions

- Incrementally pulled data to be pushed to SCD data warehouse is critical task

### Need to design central data warehouse

- In order to achieve consolidated reporting
- It was decided to build central data warehouse on redshift as data warehouse.
- Necessary skills to design data lake and reporting specific data warehouse data marts and to achieve it in required time was a challenge

## Volume of data

- Total volume of data is high, introduces challenges of first full data move, as well as in incremental data pull as well.
- Daily movement of data is also high, needs close automated monitoring to ensure everything is working fine.

## Setup data governance and observability

- Ensure low maintenance data platform
- Setup robust data pipelines which ensure data is always available
- Ensure best data governance practices

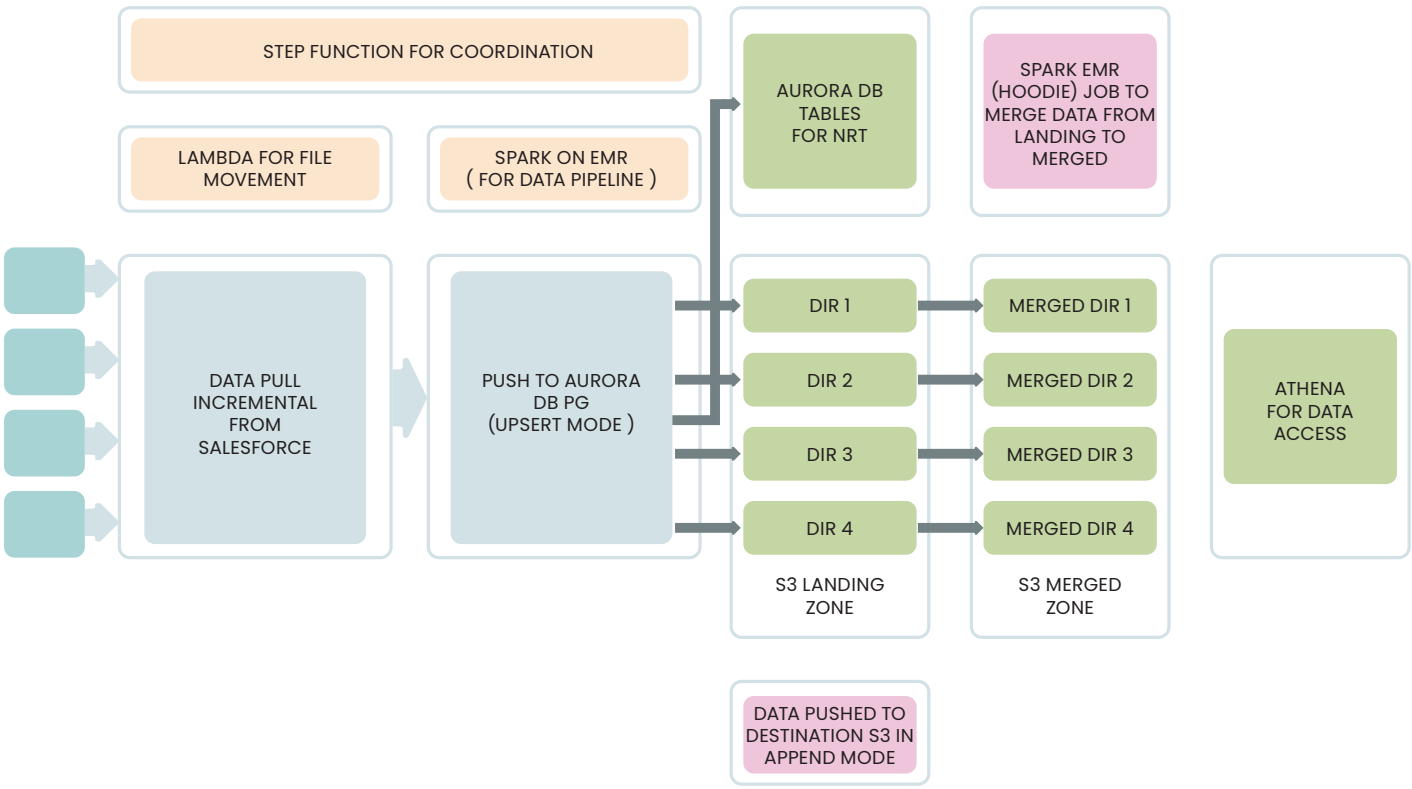
## Availability of tech experts with required skills to maintain and enhance further

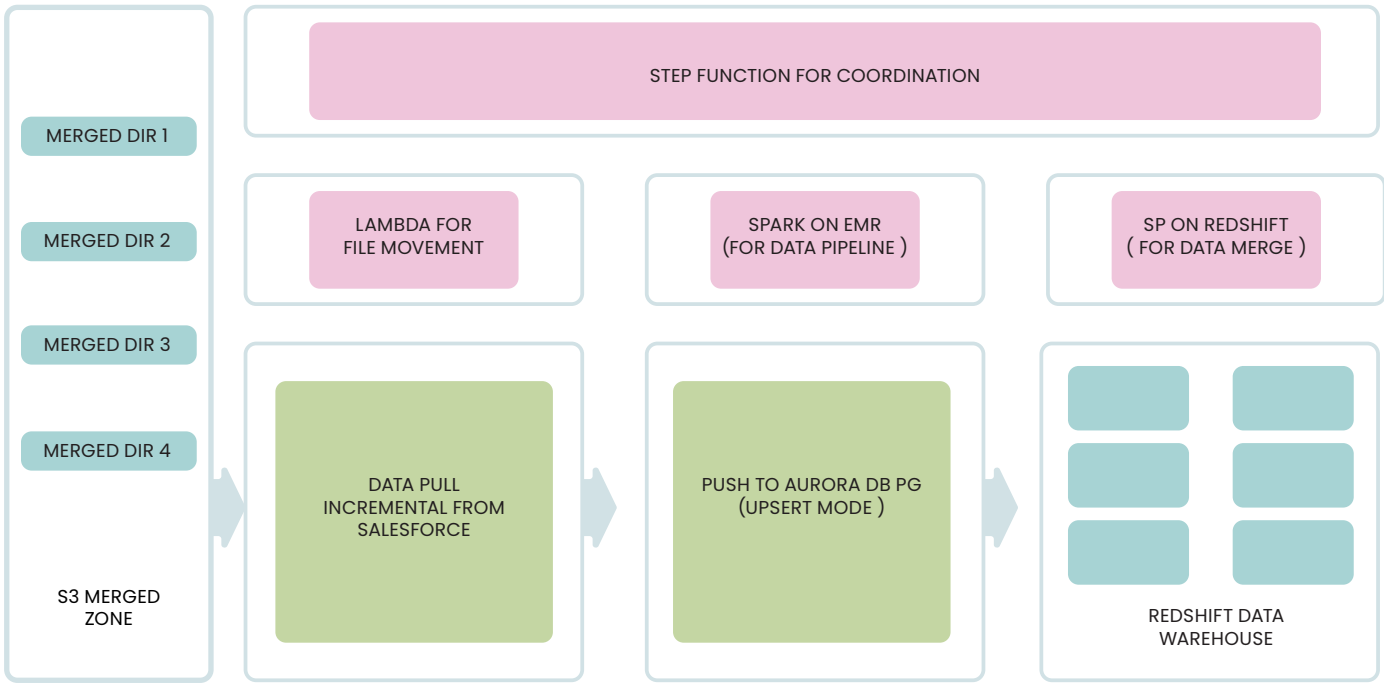
- Ensuring availability of tech experts with required skills to maintain and enhance further

# How has Dataeaze assisted the customer?

## Built modern data platform

Data onboarding to S3






## Dataeaze built a modern data platform for customer data with key features,

- A configurable data onboarding framework built on Spark on EMR
  - Incremental and efficient pull from source Salesforce
  - Efficient merge to S3 (Hudi), AuroraDB (PG), Redshift
  - Scheduled through step function and cloud watch
- Central data lake on s3 with merged layer with an access from Athena
  - Near real-time stream processing capability (With Spark on EMR)
  - Data warehouse with Redshift
  - Reporting with Tableau

## Set robust data onboarding automation

- Set continuous scheduled automation for data onboarding for
  - Source system and destination data warehouse (Based on Python and PySpark)
  - Data movement from Mysql and other sources which are backbone of some of internal systems

## Dataeaze experts getting involved to build use cases

- Dataeaze data experts are a part of the customer data engineering team, with continuous support to develop data processing pipelines for new analytics needs.

## Ensured robustness and low maintenance of automation

- Setup of Alerting of data pipelines and Monitoring framework to observe robustness of data movement

# Benefits to the Customer

## Central data platform for data across LOS, LMS, lead generation and other functions.

- Client is now enabled with the central data platform where data is getting on boarded continuously and automatically.
- Always available for analysis for analytics and data science teams.

## Achieved near real time data pull

- Some of reports require data to be refreshed within gap of 5 minutes
- Complex pipeline from maintenance perspective if there are 100+ tables. Dataeaze EDP framework achieved same.

## Easy reporting for analytics team

- The Analytics team now has a single source of data available, making it easy for reporting and analytics.
- They get robust data availability SLAs which are met with full data accuracy. Near real time and T-1 (in multiple cases) data is always available in the data warehouse.

## Robust low maintenance data automation

- Build provisioning aspects of data platform
- Alerting and monitoring to ensure data platform stability

## Accessibility to data experts

- Dataeaze data experts are involved to implement data ETL automation and pipelines
- This ensures to achieve necessary speed of development and to mitigate uncertainty of availability of experts

dataeaze

www.dataeaze.io

## About dataeaze

Dataeaze helps its customers build an analytics data platform around modern big data ecosystem. Dataeaze systems is focused on making it easy for organisations to work with data. Organisations assisted by Dataeaze get benefit of quick bring up of robust data platform with analytics capabilities brought up as per need.

contactus@dataeaze.io