

LumiQ

Prepared By:

Saksham Bairathi (2018BTECHCSE114)

Faculty Guide:

Dr. Amit Sinhal



**Department of Computer Science & Engineering
Institute of Engineering and Technology (IET)
JK Lakshmipat University, Jaipur**

May-2022

Lumiq

Practice School-2

Submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Technology in Computer Science and Engineering

By:

Saksham Bairathi (2018BTECHCSE114)

Faculty Guide:

Dr. Amit Sinhal



Department of Computer Science & Engineering

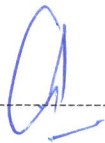
Institute of Engineering and Technology (IET)

JK Lakshmipat University, Jaipur

May-2022

CERTIFICATE

This is to certify that the Practice School-2 project work entitled **LumiQ(Crisp Analytics Pvt Ltd)**, Noida submitted by **Saksham Bairathi (2018BTechCSE114)**, towards the partial fulfillment of the requirements for the degree of **Bachelor of Technology in Computer Science and Engineering** of JK Lakshmipat University Jaipur is the record of work carried out by them under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for Practice School-II examination.



Dr. Amit Sinhal

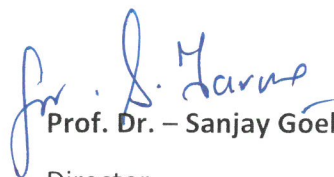
Faculty mentor and Professor,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur



Dr. Gireesh Kaushik

PS-2 Co-ordinator and Assistant Professor,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur

Date of Submission: 27-05-2022



Prof. Dr. – Sanjay Goel

Director
Institute of Engineering & Technology JK
Lakshmipat University Jaipur

ACKNOWLEDGEMENTS

The internship opportunity I had with **LumiQ(Crisp Analytics Private Ltd), Noida** was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of **LumiQ**. I am also grateful for having a chance to meet so many wonderful people and professionals who led me through this internship period.

I perceive this opportunity as a big milestone in my career development. I have taken efforts in completing this internship. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to express heartfelt gratitude to **Prof Dr.Sanjay Goel**, Director, Institute of Engineering & Technology for providing this opportunity to me. Also, I would like to thank my faculty guide **Dr. Amit Sinhal** for his constant support and guidance.

I am highly indebted to **LumiQ** for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards the members of **LumiQ** for their kind cooperation and encouragement which helped me in the completion of this project. I would like to express my gratitude and thanks to industry mentors **Mr. Harsh Choudhary, Ms. Anjali Aggarwal** for giving me such attention and time.

Yours Sincerely,
Saksham Bairathi
(2018BTechCSE114)

ABSTRACT

In any industry, analytics is given big importance in their research and development process. Analytics not only helps us analyze the facts but also boosts the decision making of any business.

Data Scientists and Analysts use data analytics techniques in their research, and businesses also use it to inform their decisions. Data analysis can help companies better understand their customers, evaluate their ad campaigns, personalize content, create content strategies, and develop products.

In the following report, I have explained how I approached and solved this problem statement. I have also mentioned all the state-of-the-art technologies/frameworks that were used in the project.

LIST OF FIGURES

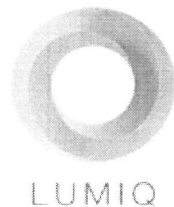
Figure	Page No.
Figure 2.4.1: Project Architecture	7
Figure 24.2: Project Architecture in detail	8
Figure 3.3.1: Send data from kafka to spark	14
Figure 3.3.2: Data from Glue to RDS	15

TABLE OF CONTENT

CERTIFICATE	i
ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
LIST OF FIGURES	iv
1. Introduction	1-2
1.1 About the organization	1
1.2 General Trainings	2
2. Project 1- Hyper Personalization	3-10
2.1 Project Objective	3
2.2 Project Structure	3-5
2.3 Technologies used	5-7
2.4 Project Architecture	7-8
2.5 Assigned tasks	8-9
2.6 Result and analysis	10
3. Project 2 - Analytical Fields Ingestion Data Lake and Personalization	11-15
3.1 Project Objective	11
3.2 Project Structure	11-13
3.3 Assigned tasks	13 -15
4. Conclusion	16
References	17

CHAPTER 1: INTRODUCTION

1.1 About the organization



LumiQ offers a wide range of data and analytics solutions. The diverse solutions portfolio is made of solutions that have the AWS technologies at their core. This makes them state-of-the-art solutions that are capable of managing and fulfilling all the data needs elegantly and efficiently.

As a data analytics company, the company's vision is to leverage data responsibly to build a safer, smarter, and more efficient solution.

LumiQ's cloud-native AI data platform enhances the customer experience by 37% and business productivity by 23%.

They are singularly focused on helping Financial services companies become analytically fluent, and help you discern which data to focus on and which to filter out so the right insights can be monetized.

Their state-of-the-art platform and solutions enhance agility and customer experience, consequently helping to differentiate and to elevate their customer's brand.

So, the ultimate objective is a human one: to enhance the quality of human life.

Website	https://lumiq.ai/
Headquarters	LumiQ, Tower-A, Noida One, Noida, Uttar Pradesh
Type	Private
Founded	2013
Specialties	Extensible Data Models, E2E AI/ML Lifecycle Management, Security and Compliance

1.2 General Trainings

- **Java** - Focused the training more on concepts like Inheritance, Interface, Constructors, Regex Functions. Given a set of questions and in addition to it some hard level hackerrank questions to understand and solve problem solving level of questions using the above mentioned concepts and Some Data Structures (Queue, Stack, ArrayList, HashSet)
- **Python** - Focused the training more on concepts like Inheritance, Interface, Constructors, Regex Functions. Given a set of questions and in addition to it some hard level hackerrank questions to understand and solve problem solving level of questions using the above mentioned concepts and Some Data Structures.
- **API** - Used Nodejs to create an API to perform crud operations, added some validations on the format of columns like email, date, etc. Used Postman to run the API.
- **Git** - Focused on some basic concepts of creating and initializing a new repository and then did git add, commit, push, pull etc. Created conflicts and solved them. Created new branches, merged them. Revert to previous commits etc.
- **JavaScript** - Basic JavaScript concepts, then made a function by scraping data from Cricinfo and visualizing it on the webpage. Made a working Calculator, and a Weather APP which could tell the weather of any place you search for.
- **Docker** - Learned basics like build, pull, push, save. Hosted a webpage using nginx and Docker on the localhost port. Connect to docker hub and push the docker image into it.
- **Cloud** - Created a AWS Lambda Function, which had an API as an trigger from which the function would retrieve message from the API and send it to DynamoDB and create an SNS notification.

CHAPTER 2: PROJECT 1 - Hyper Personalization

2.1 Project Objective

The objective of our project was to strengthen data Democratization : Access to consolidated data for users enabling data exploration for newer & deeper insights. Unify technology landscape with a single platform for many types of data workloads, eliminating the need for different services and infrastructures. Reference Architecture to support long term data strategy. Remove Duplication by consolidating ingestion pipelines, optimized storage options , removing redundant storage & redundant pipelines and creating lifecycle policies.

2.2 Project Structure

- **Data Collection (Technologies - AWS DMS, Kinesis)**

The data is collected by importing the database using AWS DMS and then the data is converted into a live stream data using AWS Kinesis the data generated is then used for the below mentioned operations

AWS DMS

AWS Database Migration Service (AWS DMS) helps you migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from the most widely used commercial and open-source databases.

AWS Database Migration Service supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle or Microsoft SQL Server to Amazon Aurora. With AWS Database Migration Service, you can also continuously replicate data with low latency from any supported source to any supported target.

AWS Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information. Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application. With Amazon Kinesis, you can ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning,

analytics, and other applications. Amazon Kinesis enables you to process and analyze data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin.

- **Data pipeline Testing**

A data pipeline is a set of tools and processes used to automate the movement and transformation of data between a source system and a target repository.

- **Data Reconciliation**

Data reconciliation (DR) is a term typically used to describe a verification phase during a data migration where the target data is compared against original source data to ensure that the migration architecture has transferred the data correctly. Data reconciliation (DR) is a term typically used to describe a verification phase during a data migration where the target data is compared against original source data to ensure that the migration architecture has transferred the data correctly.

- **Data Processing (Technologies - AWS Services)**

It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing.

- **Mapping**

Data mapping is the process of connecting a data field from one source to a data field in another source. This reduces the potential for errors, helps standardize your data, and makes it easier to understand your data by correlating it, for example, with identities.

Data mapping helps you establish a single source of truth for business-critical personal and sensitive data about your customers. It can also provide visibility into how you know those things—what actual data records you hold about your customers, which systems hold the records, and how those records are related and connected. Understanding this data at a granular level enables you to achieve deeper insight into customer preferences and behaviors.

- **Data Consumption (Technologies Spark , AWS Services)**

PySpark is a Python interface for Apache Spark that allows you to tame Big Data by combining the simplicity of Python with the power of Apache Spark. As we know Spark is built on Hadoop/HDFS and is mainly written in Scala, a functional programming language akin to Java.

- **Data Quality**
 - Null
 - format
 - Data Type
 - mismatches

2.3 Technologies Used

We have used Pyspark,SQL and AWS Technologies in making this project as mentioned below-

- **S3**

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere.
- **Athena**

Athena is an interactive query service which allows you to query tables from a database and perform operations on them.
- **Lambda**

Lambda is a compute service that lets you run code without provisioning or managing servers. Lambda runs your code on a high-availability compute infrastructure and performs all of the administration of the compute resources, including server and operating system maintenance, capacity provisioning and automatic scaling, code monitoring and logging. With Lambda, you can run code for virtually any type of application or backend service.

- **CloudWatch Logs**

CloudWatch Logs enables you to centralize the logs from all of your systems, applications, and AWS services that you use, in a single, highly scalable service. You can then easily view them, search them for specific error codes or patterns, filter them based on specific fields, or archive them securely for future analysis. CloudWatch Logs enables you to see all of your logs, regardless of their source, as a single and consistent flow of events ordered by time, and you can query them and sort them based on other dimensions, group them by specific fields, create custom computations with a powerful query language, and visualize log data in dashboards.

It allows you to monitor your lambda functions, and also allows us to create alarms in case of any errors.

- **DynamoDB**

It is a NoSQL database which allows us to store the data in key and a value pair. DynamoDB is an Amazon Web Services database system that supports data structures and key-valued cloud services. It allows users the benefit of auto-scaling, in-memory caching, backup and restore options for all their internet-scale applications using DynamoDB.

- **EMR**

Amazon EMR is a cloud big data platform for running large-scale distributed data processing jobs, interactive SQL queries, and machine learning (ML) applications using open-source analytics frameworks such as Apache Spark, Apache Hive, and Presto.

Allows to run different clusters on shell using ssh, for instance, pyspark.

- **SNS**

Allows you to send emails and notifications using the lambda function. Amazon Simple Notification Service (Amazon SNS) is a fully managed messaging service for both application-to-application (A2A) and application-to-person (A2P) communication.

The A2A pub/sub functionality provides topics for high-throughput, push-based, many-to-many messaging between distributed systems, microservices, and event-driven serverless applications. Using Amazon SNS topics, your publisher systems can fanout messages to a large number of subscriber systems, including Amazon SQS queues, AWS Lambda functions, HTTPS endpoints, and Amazon Kinesis Data Firehose, for

parallel processing. The A2P functionality enables you to send messages to users at scale via SMS, mobile push, and email.

- **SES**

Amazon Simple Email Service (SES) is a cost-effective, flexible, and scalable email service that enables developers to send mail from within any application. You can configure Amazon SES quickly to support several email use cases, including transactional, marketing, or mass email communications

2.4 Project Architecture

The raw data is imported in the AWS using AWS DMS in the form of a database. The database is then converted into a live stream data using AWS Kinesis, so the data is converted into live-stream data so that it can be accessed by other AWS services like AWS Lambda, AWS dynamodb, AWS S3 bucket, AWS Athena etc. The data then gets stored in the S3 from which the data goes into several stages like data cleaning pre-processing and data transformation etc, the transform data is then stored in a database for example - AWS RDS which is a SQL database or AWS dynamodb which is a NoSQL database. Now since the data is stored in the database we can run some queries over it in AWS Athena to get some insights from the data. Since the data is stored in the S3 bucket, which can also be used by AWS Lambda to filter out or perform some logical operations on the data and edit the data inside the database.

For security of the data we have created some IAM policies and IAM roles wherein we have provided the users who can have access to the data.

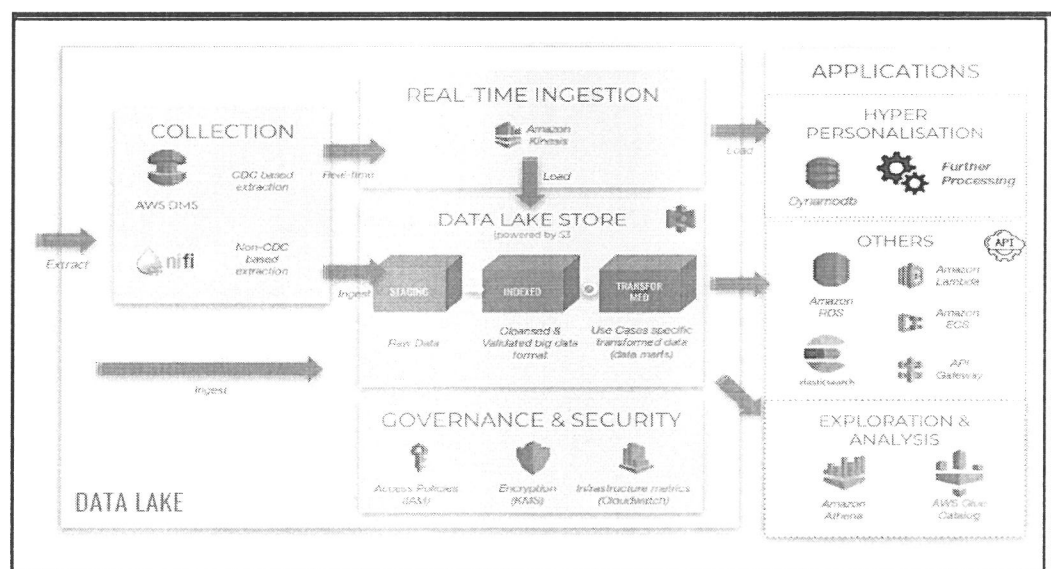


Figure 2.4.1:Project architecture

To get some quick insights or filter out garbage or null values from the data we have used spark, spark can query over the data containing millions of rows in few seconds and also unlike map reduce spark is able to run for live streaming data as well. We have used the parquet format for giving the files as an input to the spark dataframe. Also we have used pyspark for the above mentioned operations

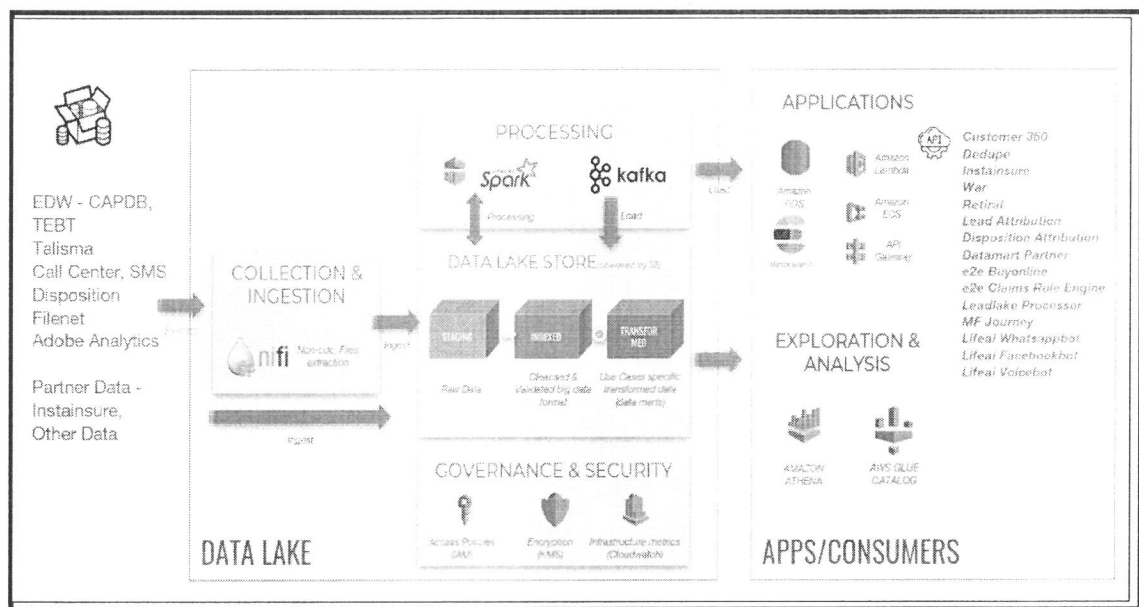


Figure 2.4.2:Project architecture detailed

2.5 Assigned tasks

1. **Errors and Exceptions in Lambdas:** In this task I was told to find the type of exceptions and errors caused for the data generated while the lambda functions are triggered and validate and let them know about the changes in the function wherever required, this data updates after every 12 hours everyday. I used to check that in CloudWatch Logs and Lambda.
2. **Number of Exceptions and Errors:** Therefore in addition I was told to make a spreadsheet containing the functions and the exceptions they are facing and what could be the reason behind those exceptions and errors if any.

3. **Not updated tables:** So the data generated through lambda functions used to get stored as a csv file S3 bucket and could be accessed in AWS Athena by some SQL queries. Now, I was told to find the tables which were not modified within a few days to have a look at them and update them in Spreadsheet. I did it by running some SQL queries on Athena. I also checked the Lambda functions from which they were getting updated from.
4. **Lambda Function:** Got a task to create a lambda function which gets triggered when a json file is inserted inside the S3 bucket. After that, verify the data and send SNS notifications and create a CloudWatch alarm. So basically what happens is that the lambda functions take the data from a json file and convert it into a parquet file. Now the thing is the json file is itself generated from another lambda function, so if the file is empty(containing no key value pairs) the lambda throws an exception. So I had to figure out the reasons for json file being empty and solve that in lambda.
5. **Spark Dataframes:** Some of the csv files contained null, NaN or garbage values so I was told to query those files by converting them into parquet and then creating a pyspark dataframe to find such values and make a spreadsheet of the primary key ids of such records. So, Did analysis by creating Pyspark dataframes and performed some filter operations using pyspark and sql. Accessed pyspark using EMR.
6. **Data Types:** Some columns had multiple data types in them, wrote a function to check their count and updated the sheet.
7. **Finding Reasons:** Dataframes had differences even though they were coming from similar lambda functions, although every column is generated from different files and lambda functions so had to figure out the names of the columns and find the reasons for their mismatches.
8. **Pyspark Function:** Created a pyspark function which finds the mismatches in the rows for the two dataframes given as an input, it also finds the null values present in the dataframes. Furthermore, it also finds the missing leads in both the dataframes.
9. **Updated Lambda Functions:** Updated some lambda functions which caused the mismatches or null values in the two dataframes and solved the issue.
10. **Docker File:** Wrote a function which finds and downloads all the files present inside a https directory. Create a docker image which consists of the virtual environment of Python. So the image could be run in any system no need to install python manually.

2.6 Results and analysis

The Data migrated into AWS using AWS DMS might contain null values or some garbage values which may be a problem for create an obstacle for further operations on the data set, the example below shows some null values present in the records of the data present in the database.

policy_number	application_login_date	application_login_date	updated_dttm	updated_dttm
24658513	2022-02-16T10:07:...	2022-02-16T10:07:25Z	2022-02-26T05:07:...	2022-02-26T06:00:...
24257714	2021-10-07T17:01:50Z	2021-10-07T17:01:50Z	2022-02-16T09:14:...	2022-02-14T03:09:...
24584471	2022-01-24T13:53:...	2022-01-24T13:53:17Z	2022-02-26T05:05:...	2022-02-11T05:13:...
24258876	2021-10-07T12:32:05Z	2021-10-07T12:32:05Z	2022-02-16T09:08:...	2022-02-14T03:31:...
23849718	2021-06-09T20:22:13Z	2021-06-09T20:22:13Z	2022-02-16T09:08:...	2022-02-14T03:26:...
22627924	2020-06-03T18:55:34Z	2020-06-03T18:55:34Z	2022-02-16T09:06:...	2022-02-11T05:26:...
17272826	null	null	2022-02-16T09:02:...	2022-02-19T21:36:...
23969319	2021-07-08T16:00:51Z	2021-07-08T16:00:51Z	2022-02-16T09:09:...	2022-02-14T04:25:...
17796182	2015-08-05T13:49:...	2015-08-05T13:49:...	2022-02-16T09:02:...	2022-02-13T03:31:...
18499894	2016-06-07T18:44:...	2016-06-07T18:44:...	2022-02-16T09:02:...	2022-02-12T22:08:...
19003187	2017-02-09T16:44:...	2017-02-09T16:44:...	2022-02-16T09:11:...	2022-02-19T00:30:...
21151687	2019-02-07T16:01:...	2019-02-07T16:01:...	2022-02-16T09:05:...	2022-02-11T04:56:...
22483084	2020-04-03T11:26:...	2020-04-03T11:26:...	2022-02-16T09:06:...	2022-02-10T04:03:...
24547146	2022-01-08T11:56:25Z	2022-01-08T11:56:25Z	2022-02-16T09:11:...	2022-02-14T02:00:...
24645886	2022-01-31T18:15:...	2022-01-31T18:15:41Z	2022-02-17T05:47:...	2022-02-16T05:24:...
23968863	2021-07-06T11:31:08Z	2021-07-06T11:31:08Z	2022-02-16T09:07:...	2022-02-14T03:28:...
16956775	null	null	2022-02-16T09:09:...	2022-02-14T02:33:...
17477885	null	null	2022-02-16T09:09:...	2022-02-14T03:54:...
18672243	2016-09-03T11:39:...	2016-09-03T11:39:...	2022-02-17T05:39:...	2022-02-16T04:15:...
18999242	2017-02-08T12:46:...	2017-02-08T12:46:...	2022-02-16T09:02:...	2022-02-12T01:51:...

As we can see above there are some rows where in the column contains null values, so basically this data is generated by some python functions, and in order to figure this out that why this column is null for these particular policy_number we need to see the logic in that python function and figure out why there is null value and correct the logic and rerun the function which will solve this issue right away.

Similarly there were some more columns like these which contained some garbage values and those values can be fixed by visiting the respective lambda function and fix the logic and rerun the function.

CHAPTER 3: PROJECT 2 - Analytical Fields Ingestion Data Lake and Personalization

3.1 Project Objective

The aim of the project is to provide a single source of service for different needs of the client and also to provide a cost efficient solution to the problem, the client has the data generated in the very large quantity and wants to get insights from the data , so they want the data to be transformed and profiled to have better understanding of the data.

3.2 Project Structure

- **Data Collection (Technologies - AWS DMS, Kinesis)**

The data is collected by importing the database using AWS DMS and then the data is converted into a live stream data using AWS Kinesis the data generated is then used for the below mentioned operations

AWS DMS

AWS Database Migration Service (AWS DMS) helps you migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from the most widely used commercial and open-source databases.

AWS Database Migration Service supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle or Microsoft SQL Server to Amazon Aurora. With AWS Database Migration Service, you can also continuously replicate data with low latency from any supported source to any supported target.

AWS Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information. Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application. With Amazon Kinesis, you can ingest real-time data such as video, audio, application logs,

website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications. Amazon Kinesis enables you to process and analyze data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin.

- **Data pipeline Testing**

A data pipeline is a set of tools and processes used to automate the movement and transformation of data between a source system and a target repository.

- **Data Reconciliation**

Data reconciliation (DR) is a term typically used to describe a verification phase during a data migration where the target data is compared against original source data to ensure that the migration architecture has transferred the data correctly. Data reconciliation (DR) is a term typically used to describe a verification phase during a data migration where the target data is compared against original source data to ensure that the migration architecture has transferred the data correctly.

- **Data Processing (Technologies - AWS Services)**

Data Processing includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing.

- **Mapping**

Data mapping is the process of connecting a data field from one source to a data field in another source. This reduces the potential for errors, helps standardize your data, and makes it easier to understand your data by correlating it, for example, with identities.

Data mapping helps you establish a single source of truth for business-critical personal and sensitive data about your customers. It can also provide visibility into how you know those things—what actual data records you hold about your customers, which systems hold the records, and how those records are related and connected. Understanding this data at a granular

level enables you to achieve deeper insight into customer preferences and behaviors.

- **Data Consumption (Technologies Spark , AWS Services)**

Spark

PySpark is a Python interface for Apache Spark that allows you to tame Big Data by combining the simplicity of Python with the power of Apache Spark. As we know Spark is built on Hadoop/HDFS and is mainly written in Scala, a functional programming language akin to Java.

- **Data Quality**
 - Null
 - format
 - Data Type
 - mismatches

3.3 Assigned Tasks

- **Data Profiling -**

Data profiling is a process of reviewing and analyzing diverse datasets across the business to inform business decisions. As your business grows and evolves, it will generate large amounts of data around customer purchase history, business spending history, accounting and finance, operating metrics and more. Without data profiling, some potentially useful and valuable data could get pushed to the back of the virtual filing cabinet, out of sight and out of mind, and its potential value is lost. By using data profiling, your business creates a useful, searchable inventory of your business data. This can help create better and more useful business intelligence reports, improve efficiency and costs, and lead to better profits in the long term. When data profiling, your team or systems intelligently examine existing data to understand the data structures and any potentially related data.

So the data contained some attributes which either contained null values , garbage values or values not in the correct format. And some unwanted columns as well. So I was provided with a schema and I was asked to convert the csv file into this particular schema and also to put some validations so that the data comes in the correct format, I was asked to use scala for this.

The steps I followed to complete the same were

- Wrote some SQL queries to validate the columns like (driving license, pan card, aadhar etc) into correct format.

- Integrated the SQL queries in Scala and using Struct, I converted the data into the desired dataframe.
- The AWS service used in this was EMR to access spark, and later I wrote the final dataframe into a csv format in S3 bucket.

- **Kafka -**

Kafka is primarily used to build real-time streaming data pipelines and applications that adapt to the data streams. It combines messaging, storage, and stream processing to allow storage and analysis of both historical and real-time data.

So there were some JSON files, which were being created daily, and then were imported in spark as a dataframe to perform some queries, therefore in order to automate this task I was asked to write a glue job which would send the json files using kafka to spark dataframe and perform some queries over it.

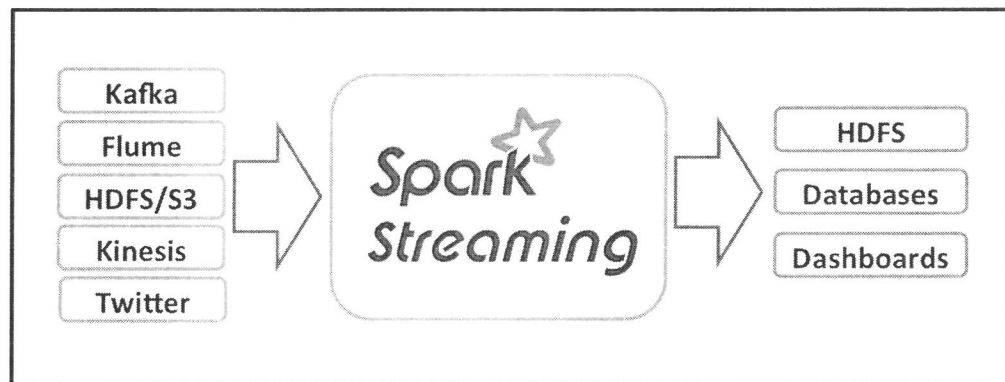


Figure 3.3.1: Send data from kafka to spark

The steps followed were -

- Setup kafka in a EC2 instance which is setup and the brokers provided are with AWS MSK
- Wrote a function which gets the generated S3 files and sends them using kafka producers. Using the brokers provided by the MSK
- The file is then retrieved in the pyspark dataframe using kafka consumer.
- And as the set of spark queries are already defined in the function, the function will perform them and save the file in the S3.

- **AWS RDS -**

Amazon Relational Database Service (RDS) is a managed SQL database service provided by Amazon Web Services (AWS). Amazon RDS supports an array of

database engines to store and organize data. It also helps with relational database management tasks, such as data migration, backup, recovery and patching.

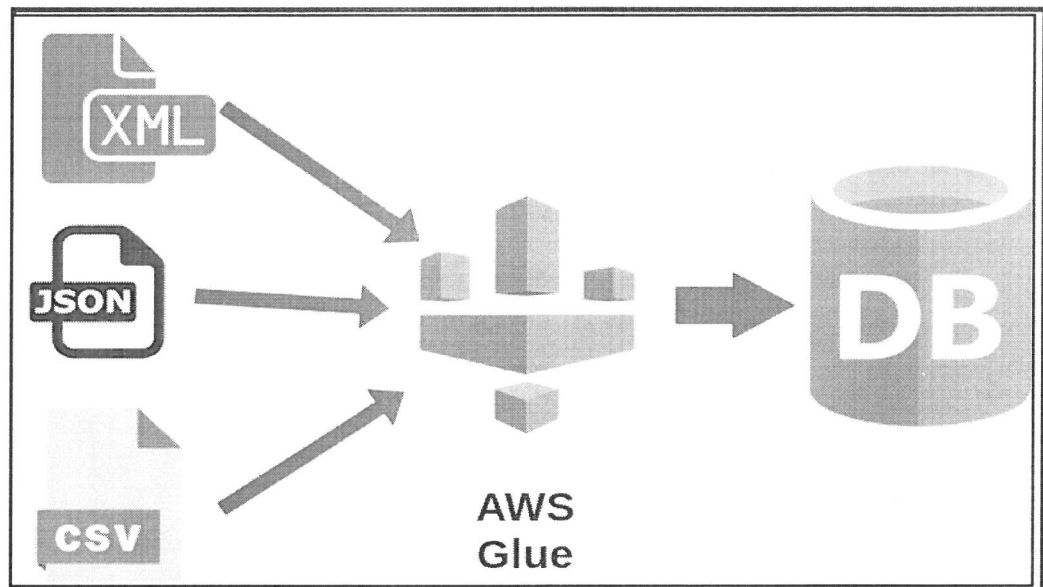


Figure 1.3.2: Data from Glue to RDS

Amazon RDS facilitates the deployment and maintenance of relational databases in the cloud. A cloud administrator uses Amazon RDS to set up, operate, manage and scale a relational instance of a cloud database. Amazon RDS is not itself a database; it is a service used to manage relational databases.

So the task was to insert the data present in the csv file to the RDS Database, the steps I followed to get the task done were.

As shown in the image above.

- Wrote a lambda function to get the S3 location from lambda to Glue.
- After retrieving the files in glue from the S3 location.
- I used pyspark to convert the csv files into spark data frame
- Later I created a RDS Aurora Postgres Database.
- And using JDBC I set the connection with the created database.
- And wrote the data from spark dataframe to RDS.

CHAPTER 4: CONCLUSION

At Crisp Analytics Pvt. Ltd. The Practice School- II opportunity provided a lot of knowledge in advanced Big Data technologies like Spark, SQL and AWS Services. I can affirm that the professional experience and knowledge I acquired here would really assist my overall career. I have accomplished my 5-month internship learning lots of new tools and technologies. Working in a professional setting, such as Crisp Analytics Pvt. Ltd, has also taught me valuable skills such as teamwork and basic professional work ethics. The mentors here are extremely helpful and supportive. Moving on to the project work. My contributions to the project work at the company were highly appreciated by my mentors and my seniors. Working on a real-life Big Data project has improved my knowledge and skill sets. During this internship period, I have concluded one major project named “Hyper Personalization” and have started working on another one.

REFERENCES

1. AWS Lambda Services: <https://www.youtube.com/c/AWSTrainingCenter>
2. Pyspark course - <https://www.youtube.com/c/Freecodecamp>
3. Docker Course -<https://www.youtube.com/c/CodeWithHarry>
4. Kafka course -<https://www.youtube.com/c/StephaneMaarek>
5. Scala Course - <https://www.youtube.com/c/ProgrammingKnowledge>

INTERNSHIP CERTIFICATE

Date: 20/05/2022

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Saksham Bairathi has done his internship in Big Data at Lumiq.ai, Noida, from 3rd Jan' 2022 to 20th May 2022 for the duration of 4.5 months.

He has worked on a project titled Hyper Personalization. This project was aimed to personalize and recommend agents based on lead and agent data. As part of the project, he has worked on data profiling, data cleaning and root cause analysis.

During his internship he has demonstrated his skills with self-motivation to learn new skills. His performance exceeded our expectations, and he was able to complete the project on time.

We wish him all the best for his upcoming career.

Lumiq.ai

Sneha Pandey

HR

[Signature][Seal of Company]

