**Syed Hussain ** 

BigData Developer

London, UK | +44-7498969898 | [syedprogramming@gmail.com](mailto:syedprogramming@gmail.com) |

Hacker Rank: @ syedprogramming |

Git Hub: https://github.com/SyedProgramming

* Worked with analysts and finalize the requirement document. As part of requirement, get the source table DDLs from client.



**PROFESSIONAL EXPERIENCE**

* AES256 Key Encryption SDS (Secure Data Service) framework development to process 100 million records in a minute using Distributed Programming.
* Generate the Avro schema files from input source DDL got as part of requirement.
* Active-Active Data Centers.
* As per need, develop the spring xd module and test the module by launching sample spring xd jobs and pull the data.
* Deploy the project and test end to end on Dev and Load test environments.
* Identify the PII/ PCI fields and apply scrubbing rules as per client requirement for data security.
* As part of initial analysis, decide what should be fetch size of each Spring XD job and create job statements.
* Decide the Kafka topics to be created, partitions under each topic, replication factor needed and create topics accordingly. Topics will be created on both networks. (Client network and IBM SoftLayer).
* Create the Kafka mirror for each source system, decide and allocate the appropriate disk space for the mirror as per the data.
* Create the Hive external and internal (ORC) tables to hold the data.
* Analyze the count and data present in Hive tables so that will be available to end user data scientists.
* Use Hive joins to analyze the data between tables.
* Create the Camus property files as per the data source and do throughput analysis for number of days Camus will take for data processing.
* As per the Camus throughput change the Kafka topic retention periods and partitions to avoid data loss and process data within time.
* Validate Avro formatted data on Hadoop/ Hive immutable once Camus processing is done.
* Move the data to Hive ORC tables with another utility called IDS Ingester.
* Validate the ORC formatted data, which is consumed by data scientists later.
* Handle the cases where the schema changes will be there by creating temp Hive tables and moving data later to permanent tables.
* Check the status of spring xd jobs and analyze the logs on Splunk.
* Create the Splunk alerts for errors of Camus, Spring XD Jobs, Hive etc.
* Worked with Spark, Scala and Kafka steaming.
* Handled different file formats like JSON, XML, Fixed Width file, Avro, Flat file during data pipeline and POCs in Hadoop and Hive.
* Carried out historical load activity via Sqoop.
* Write Sqoop queries and jobs to pull the data from MySQL, Oracle etc. sources and put it under HDFS.
* Write Sqoop queries to load data under Hive tables.
* Write Hive and Impala queries to do sanity testing of data loaded under IBM SoftLayer.
* Monitor and check status of Sqoop jobs from Cloudera Manager.
* As per the client requirement handle PII/PCI fields from tablesaccordingly.
* Write shell scripts to load the data in JSON and XML format to HDFS and on top of that create Hive tables and verify data under it.
* Write shell scripts to preprocess and load Fixed Width files data to HDFS. Once the data is loaded to Hadoop create Hive tables on top of it.

**SUMMARY**

**Over 8+ years** of hands on experience in design, development and delivery of robust, scale-able solutions.

**BigData Developer** with proven ability to meet high-pressure deadlines and coordinate multiple teams, projects.

**Backed by strong industry proficiency** including mission critical releases of highly available systems, Agile methodologies; as well as cross-functional skills in Product Management.

**Proven prowess** in handling large production stacks by creating scale- able system architectures while keeping a check operational costs.



**SKILLS**

**JAVA**, **Scala**, **Python** **Spark**, **Kafka**, Spring Streams, Netty, Tomcat, WebSphere

HBase, MySQL, Oracle, Cassandra HDFS, Hive, Pig, Flume, Oozie Sqoop, Zookeeper

Struts, Hibernate

ElasticSearch, Logstash, Kibana Amazon Web Services (AWS) ReST, WSDL, Webservices



JDBC,HTML, CSS, JQuery, Ajax,

XMl Parsers – Jax-B, Jax-P, SAX, JSON Transfer,Swings, Applets **Docker**, **Kubernetes**, Jenkins, Jira, Git, Ansible, Maven, Ant, SBT, Gradle, Grafana, DataDog, **AirFlow**, Jacoco - Spark Unit Test Cases, Sonarqube Code Coverage.



**EDUCATION**

**Osmania University** Master of Computer Science 2011

 82%



**April’21 – Till Date Data Engineer VCA – Visa Europe**

VCA Datamart performance tuning and optimization:

Worked on optimizing the Spark code which was impacted our delivery. Tried the workaround with different performance tuning options and by applying multiple hyperparameters. Able to complete the query in 4 hours which was taking 72 hours earlier.

Optimized Pyspark code to meet the Design Patterns and oops concepts in a functional programming way. This demonstrates "We Lead By Example"

Automation for Barclaycard Rewards using Tuber:

Created a Tuber workflow for the Barclaycard rewards program, which helps internal stakeholders to straightforward push the file via SFTP operator. As SFTP doesn't support Tuber so this process we are unable to make it on Tuber. We can come up with a user-defined SFTP operator by extending the airflow classes but that will be VISA compliance for SFTP. Collaborated with multiple Data Scientists (Martin, Gianluca, Marius, Chris) to understand the requirement for achieving the proper automation. This demonstrated how "We Collaborate".

Involved in various activities in DPI Migration:

As part of a Data Engineering team, it's our role and responsibility to support the Data Scientist for the easy migration to the DPI environment from EPI. Helping others on multiple issues and providing workarounds on different stacks like Hive/Tez, Presto, Tableau, Spark, Tuber. Provided training to the Data Scientists, approaching them to help, logging issues, and providing support.

Raised multiple defects on tools on the DPI Prod cluster:

Provided support to the Data Platform team by identifying bugs on the tools like Port of Entry, Wrappers and also helped them to resolve the issues. Helped the Tuber team with multiple issues and provided a workaround. Highlighted certain issues related to clusters like a queue, application manager, etc, and worked with the DataPlatform team for the resolution. Provided Port of Entry, Wrappers Tool demonstration to the users of Europe region and helping them offline to use the utility without any issues.

PAN Proxy tool to perform Encryption/Decryption operations:

Helping various Data scientists with the PAN-related data, on the request basis by performing Encryption/Decryption using Pan Proxy utility. This process helps to expedite delivery with the clients. Helped Ali Yagmur to understand the Pan Proxy functionality and provided more insights on SDS and its internal execution patterns. This demonstrated how "We Collaborate"

​

Automated auditing functionality for VCA Europe Datamart:

a. Placed an audit functionality in an automated way for the VCA DSL Datamart, which helps us to identify the issues at runtime and quickly to apply the resolution. As VCA Europe DSL Datamart is a daily load, need to perform audits check daily in an automated way, implemented the functionality using Tuber.

b. Demonstrated collaboration and believed in open communication, when the team was facing performance challenges in the project. Highlighted the performance challenges in the open forum of TRC where we received valuable suggestions such as an increase in VCores, one-time

spark UDF initialization, use hive on spark engine, etc.

c. Followed up with infra-team to get the required resource in place and worked with GDS Data Engineering team to implement the suggestion received.

d. This helped in removing the barriers for the application and let application development keep going smoothly. This demonstrated how "We Collaborate".

Trained on implementing 5/50 rule with full compliance:

Properly took training on the 5/50 rule which helps me to provide solutions to the business requirement as per the VISA compliance.

Facilitator of Visa University for Spark, Hive, and Kafka tech stack:

Working as a Facilitator of Visa University for providing training for the modules Spark, Hive and Kafka. Had a couple of meeting with the content providers i.e. Elephantscale and worked on multiple solutions to make the content more useful and helpful to the learners. Had a couple of meeting with Commercial Colleges, Visa University people Eduardo Tamayo for the creation of content. Created lab examples in the MLP platform so that those who are new to MLP can explore PySpark in MLP. This demonstrated "Taking the Next Step in your Career"

Support the development of scaled solutions within Service Lines by ensuring base data:

Providing solutions to the service lines and helping them to execute their queries without any defect, which is helping them to deliver the projects on time.

Showed passion and perseverance while working on architecture, design, development of Visanet migration project using latest spark 2.2 version.

Though faced a lot of challenges while coding using spark 2.2, it benefited in writing with minimal lines of code and with enriched spark APIs.

This will help the application to be more stable in the future and would be able to integrate easily other tools using the latest API.

Requirements are streamlined and optimized through standard datamart initiatives:

Each and every requirement we are tracking on Task under Teams channel and in on Daily Standup going through the task list to track

1. What we have done yesterday

2. What we are going to be doing today

3. Raise defect removal efficiency

4. Reduce bad-fix injections

5. Eliminate security flaws in all software applications.

6. Reduce the odds of cyber-attacks

7. Lower maintenance and warranty repair

8. Improve the volume of certified reusable materials

9. 100% SDS integration for PAN encryption /decryption

10. Improve the volume of certified reusable materials

11. Improve average development productivity function points per month

12. Improve work hours per function point

13. Engineering Best Practice & Innovation

14. 24\*7 Operations available by zero downtime

15. Criticality identifies the level of risk associated with loss of servers

**Nov’17 – Till Date, BigData Developer Visa Inc Pvt. Ltd. India**

**VCA – Visa Consulting & Analytics – CEMEA Data Engineering**

As part of CEMEA Data Engineering team we provide services to our internal stake holders Data Scientist team as well as direct clients.

We provide solutions to Russia, MENA, SSA regions. We have developed a CEMEA Datamart SSOT which is the base data lake for every  
 region.

* Create and maintain optimal data pipeline architecture.
* Assemble large, complex data sets that meet functional / non-functional business requirements.
* Identify, design, and implement internal process improvements: automating manual processes, optimizing data delivery, re-designing infrastructure for greater scalability, etc.
* Build the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using Spark-Scala, PySpark, Hive, Sql, Kafka, HBase.
* Build analytics tools that utilize the data pipeline to provide actionable insights into customer acquisition, operational efficiency and other key business performance metrics.
* Work with stakeholders including the Executive, Product, Data and Design teams to assist with data-related technical issues and support their data infrastructure needs.
* Project development using object-oriented/object function scripting languages: Python, Java, Scala, Shell Script etc.
* Automation deployment using Jenkins, Tuber.
* Develop CI/CD principles. Develop and maintain pipeline configurations
* Implement and improve monitoring and alerting. Build and maintain highly available systems on Kubernetes.
* Implement an auto-scaling system for our Docker on Kubernetes nodes.

# D DTPS – Debit Transaction Processing System – DataPlatform

* + Get the transactions data from the Tandem Systems using Shell Script’s for every 10 minutes.
  + Using Spark Stream with Kafka Util load the data into topic after performing enrichments to the fields.
  + Created a Parser to read fixed length file and convert into delimited format.
  + Perform Dedup Operations for every sequence, as it is a Low Latency applications we get data for every 10 minutes per 144 sequences/cuts.
  + Each sequence/cut contains 1.5 million of data.
  + Load enriched data into Hive Table.
  + Finally push the data from Hadoop to DB2 as Micro strategy team need data into DB2 to generate reports.
  + Migrate Ab-Initio graphs to Hadoop using Hive and Spark.
  + Understand Ab-Initio graph and come up with the design in Spark or Hive based on the criticality of the graph.
  + Created an generic Utility to push the data into DB2 using Map Reduce.
  + Handle Dual Data Center’s (Active – Active Data Center).
  + Push the code to Git.
  + Build the Code using Jenkins, internally added plugins Sonar Cube for code coverage, Check Marx for Code Vulnerabilities, Maven for the Build.
  + Following Agile Methodology, daily Stand-up’s and scrum calls.
  + Jira Update every day.
  + Creation of Docker Images for an installation – POC is in progress for DevOps
  + Architectural Design and creation of Flow & UML diagrams.
  + Code review.
  + Spark Unit Test Cases.
  + Data Validation using RFAM Utility across data centers.



**Jan’15 – Oct’17, BigData Developer IBM India Pvt. Ltd. India**

# GEICO – CCIS DIVA Project

The Government Employees Insurance Company (GEICO) is an American auto insurance company headquartered in Chevy Chase, Maryland. It is the second largest auto insurer in the United States, after State Farm. It is a wholly owned subsidiary of Berkshire Hathaway that as of 2015 provided coverage for more than 22 million motor vehicles owned by more than 14 million policy holders

## Data Ingestion –

This project is regarding building data pipeline to ingest the data from client network on daily basis to IBM SoftLayer cloud. To build the data pipeline, Spring XD and Apache Kafka components are used extensively along with other big data ecosystem components. As an end result, pipeline will make data available to Analysts in form of Hive tables on Data Lake present on IBM SoftLayer. Herein, ingested and handled Terabytes of data in spring xd and at Hive/ Hadoop level

* + Work with analysts and finalize the requirement document. As part of requirement get the source table DDLs from client.
  + Generate the Avro schema files from input source DDL got as part of requirement.
  + As per need, develop the spring xd module and test the module by launching sample spring xd jobs and pull the data.
  + Deploy the project and test end to end on Dev and Load test environments.
  + Identify the PII/ PCI fields and apply scrubbing rules as per client requirement for data security.
  + As part of initial analysis, decide what should be fetch size of each Spring XD job and create job statements.
  + Decide the Kafka topics to be created, partitions under each topic, replication factor needed and create topics accordingly. Topics will be created on both networks. (Client network and IBM SoftLayer).
  + Create the Kafka mirror for each source system, decide and allocate the appropriate disk space for the mirror as per the data.
  + Create the Hive external and internal (ORC) tables to hold the data.
  + Analyze the count and data present in Hive tables so that will be available to end user data scientists.
  + Use Hive joins to analyze the data between tables.
  + Create the Camus property files as per the data source and do throughput analysis for number of days Camus will take for data processing.
  + As per the Camus throughput change the Kafka topic retention periods and partitions to avoid data loss and process data within time.
  + Validate Avro formatted data on Hadoop/ Hive immutable once Camus processing is done.
  + Move the data to Hive ORC tables with another utility called IDSIngester.
  + Validate the ORC formatted data, which is consumed by data scientists later.
  + Handle the cases where the schema changes will be there by creating temp Hive tables and moving data later to permanent tables.
  + Check the status of spring xd jobs and analyze the logs on Splunk.
  + Create the Splunk alerts for errors of Camus, Spring XD Jobs, Hive etc.
  + Worked with small POCs of Spark, Scala and Kafka steaming.
  + Worked on POC for Apache NiFi to check its usability in built data pipeline.
  + Handled different file formats like JSON, XML, FixedWidth file, Avro, Flat file during data pipeline and POCs in Hadoop and Hive.

## Historical Data load –

Historical load of data to IBM SoftLayer Cloud was one time activity and carried out with CDH distribution. This activity was carried out with Sqoop mostly.

* 1. Work with analysts and finalize the requirement documents for historical load. Identify the source system database table’s needs to pull and columns per table.
  2. Carried out historical load activity via Sqoop.
  3. Write Sqoop queries and jobs to pull the data from MySQL, Oracle etc. sources and put it under HDFS.
  4. Write Sqoop queries to load data under Hive tables.
  5. Write Hive and Impala queries to do sanity testing of data loaded under IBM SoftLayer.
  6. Monitor and check status of Sqoop jobs from Cloudera Manager.
  7. As per the client requirement handle PII/PCI fields from tables accordingly.
  8. Write shell scripts to load the data in JSON and XML format to HDFS and on top of that create Hive tables and verify data under it.
  9. Write shell scripts to preprocess and load FixedWidth files data to HDFS. Once the data is loaded to Hadoop create Hive tables on top of it.
  10. Handle data especially for the tables having different number of columns in historical table and daily load active tables on Hadoop.

## Deduplication Utility –

In above data pipeline, one issue was, some tables were having duplicate records. So to resolve this issue, we developed small utility in Java which will identify duplicate rows in Hive tables and create the new Hive table with unique rows. This project proved to be highly useful and working best in the production environment.

## Avro Schema Generator –

For each data source, first part of development were starting with creating Avro schema files for each table. So we developed a small utility using Java which will create the .avsc (Avro schema) files from source DDLs. As per the source data system (MySQL/Oracle/DB2 etc) all datatypes are mapped to corresponding Hive datatypes. This project is now working as base for all new source systems to be ingested in data pipeline.

## Hive Table Create Utility -

Depending on each data source system to be pushed into Hadoop layer, developers were doing create Hive table create statements manually and this was cumbersome task. We automated this thing and created Java utility which will create the Hive create table statements from source DDL file and data source type like MySQL, Oracle, DB2 etc. With this utility Hive external and internal table create statements were generated in form of hiveql file.

**Nov’12 – Dec’14, Java Developer IBM India Pvt. Ltd. India**

**Individual Policy Inquiry – Standard Insurance**

Developed and Supported Individual Policy Inquiry application, it is used by Individual Division producers to look up policy and new business information based in the policy id and producer name. Individual Policy Inquiry application is dependent on the ID3 application for its functionality. Policy Inquiry is secured using Tivoli Access Manager and application oracle tables. For the ID3 policy system, this integration shall be provided via an existing JDBC driver within our enterprise architecture for accessing the system’s underlying DB2/AS400 data store directly from Java. The SIC Internal users uses this tool to check what are the policies are issued to the companies and who is the producer for that policy and who are the policy holders under that producers.

* Involved in giving training on Struts Development to the development team.
* Supported team members on Application development who are new to Struts, Spring Development.
* Have taken a lead role from Development team and played a key role in the development of theproject.
* Involved in developing the initial code framework of the application.
* Responsible for giving KT to the new joined resources in the team.
* Involved in refactoring and reviewing the implemented code to make it more reusable and modular.
* Learned Jquery and got a pretty good knowledge after implementing in the application.
* Apart from the regular project work I have completed OCJP6 & OCWCD Certification which helped in implementing new approaches in the application.
* Involved in design phase which includes the Sequence and Class Diagrams
* Responsible in using Struts Framework MVC for controlling central point request handling.
* Development of DAO’s and Interfaces. Coding, Development and Deployment of Struts Action classes and UI pages on Websphere7.0.Involved in client side validations and peer reviews.
* Involved in creation of Form validations with java script
* Developed Value Objects based on schema
* Perform User Interface requirements gathering, Low Level Design (LLD) documentation, Functional Design Documentation, Build (Coding-Unit Test-Defect Fixing), CR’s (Change Requests)
* Coding, Development and Deployment of Struts Action classes and UI pages on Weblogic.
* Involved in developing Business and Transaction logic in Service layer and persistence logic in DAO layer using Hibernate.
* Writing and executing Unit Test Cases and Unit Testing.

**Environment**: Struts, ReST Webservices, Spring, Hibernate, IBM-Websphere, Oracle10g, Clear Case, and Toad, J2SDK1.5, Struts, Hibernate, Servlets2.4, Jsp2.0, Oracle9i, MyEcllipse6.5, Mock Struts Test Case, Weblogic server.

**Achievements**

1. Received a Best Employee Award in 2017 from clients in IBM India Pvt. Ltd.
2. Received a Go Beyond Award in 2020 from Vice President in VISA India Pvt Ltd.

**Certifications**

## Oracle ID: OC1230321

* + **1Z0-851** Java Standard Edition 6 Programmer **OCJP6** (Oracle Certified Java Programmer) – Score

## 98%.

* + **1Z0-858** Java Enterprise Edition 5 web Component Developer **OCWCD** (Oracle Certified Web Component Developer) – Score **100%.**

1. **NIIT** – Diploma in **Java** Technologies from National Institute of Information Technology.

## PG Diploma of Machine Learning and Deep Learning from IIIT-Bangalore(Upgrad)