**Gokulankannan Arumugam**

**Phone: +65-94663223**.

**Email: gokulit09@gmail.com**

**Objective :**

Business minded AI based Data Scientist with demonstrated ability to deliver valuable insights from Data Lake on the Advanced Big Data Analytics technology via analytics and advanced data driven methods

**Summary:-**

* **4+ years of experience in Datawarehousing Business Intelligent Platform and Spark based Big Data Development and Analysis in Advancing DBS with AI team.**
* Implemented **Spark Analytical Datamart** using Scala and Java for **Bank Loyalty Program**
* Established the DBS BankWide Enterprise DataIngestion Framework to ingest data from various source to Staging using **TALEND Tool for ETL Operations and Big Data Computations.**
* Implemented **Spark Analytical Datamart** using Scala and Java. Build the **Star schema and SnowFlake schema based OLAP Cubes** accordingly for the reports to be generated for the projects
* **Schema generation and Optimized queries** are generated in Hive on MapReduce, Hive on Spark, Impala, SparkSQL. Readily available for the loaded tables in Impala and Spark
* Enhanced **Automated Encryption of HIVE** user layer sensitive column information
* Enabled **Realtime Ingestion** on Framework using **Flume**
* Designed Hadoop Stage history Maintenance on **HIVE for Analytical DataModel.**
* Designed Framework to **Parse QlikView log** files and provided user layer on top of **IMPALA**
* Fine tune and optimized **HIVE and Impala performance**
* **Automated Loading of Transactional / Non-Transactional** table Data into HIVE from Teradata using SQOOP by reducing time latency of data loading and increased performance by handling multiple scenarios
* Designed **Java xml VTD Parser Framework** to increase the performance and reduce time latency using **SpringBoot**
* Experience in **Spark, Scala and HDFS** to manage Extraction, Transformation, Load by the action of Cleansing, Validation, Transformation and handling for SCD types.
* Designed **End-to-End Airflow Scheduling and Monitoring** Tool Server Setup, Development and **Integrated with Talend**.
* Configured the **Airflow cluster using MariaDB** as a MetaStore and **RabbitMQ** for the Scheduler Messaging on Top of 3 Clusters.
* Ingested Logs flow from Talend, Spark, Hadoop and AirFlow to **ELK**(LogStash, ElasticSearch and Kibana)
* Expertise in **Database Design**, Development and Management of Schemas, automation using **Stored Procedures, Functions, DDL, DML SQL queries.**
* Proficient in **RDBMS concepts with Hadoop Hive, Impala, Teradata, DB2.**
* Expertise in Data Warehousing & Data Integration tools and Extensive experience **in ETL (Informatica, Talend and Shell Scripting).**
* Provided Solutions, Designed and Automated **Datawarehousing SCD Types for Hive.**
* Worked on **Batch and Realtime ingestion** with various Data Source and Data platforms using **IBM** **Tivoli, Flume, JMS, Kafka and Shell Scripts.**
* Extraordinary **POC** onFinding sensitive attributes based on column level and row level analysis and build predictive model using **R.**
* POC on **NoSQL Databases** like **HBase, Accumulo, Casandra and Elastic Search.**
* Integrated Hadoop technologies as Hive, Sqoop, Linux & Unix Scripting, Java, Custom MapReducer files deployment **via Continuous Integration/Continuous Delivery CI/CD Method using Git, Bitbucket, Jenkins, Docker, Nexus, CI/RA and Artifactory**

Professional Experience:-

* Working as a “**Senior Data Engineer**” in Accision Labs, Singapore for the Client ‘**UOB**’ for the “**Power to Predict**” from Jun 2018 till date.
* Played role as a “**Junior Big Data Architect**” in **Optimum Technologies** **(P) Ltd**, **Singapore** for the Client **‘DBS’** for the **Business Intelligence Platform** from Sep 2016 to May 2018.
* Played role as a “**Big Data Developer**” in **Optimum Technologies** **(P) Ltd**, **Singapore** for the Client **‘DBS’** for the project **of ACS and Digital Banking** from Jul 2015 to Aug 2016.
* Played role as a “**Software Engineer**” in **Skava Systems**, Coimbatore from May 2014 to May 2015.

Certifications:-

* Course Completion Certification on **“Cloudera Spark and Hadoop Developer” at Cloudera on Demand Training.**
* Course Completion Certification on **“Scala” at Cloudera on Demand Training.**

Recent Automations:-

* Spark Analytical Datamart using Scala and Java for the OLAP based business Analysis.
* Encryption of Hive user layer sensitive information
* Realtime Streaming Application to load data from IBMMQ to HBASE using Kafka and Spark
* Spark ETL Pipeline to load Data from Source to Target
* Build Framework to Auto Ingest SCT Type2 and Transactional Files
* Source File Validation Framework for ingestion using Scala and Java
* ETL Computation Framework on Hadoop using Spark and Hive
* WEB HDFS and HDFS JAVA API Framework using Spring Java
* ETL and Computation Framework Deployment using Jenkins and GIT
* Integrated Airflow and Talend Ingestion Framework
* Airflow Cluster configuration and Generate Airflow DAG’s
* Established Auto Data Masking Framework in Hadoop EcoSystems
* Business Date Feature in Airflow Open Source Tool
* Automated S3 File Sensor in Airflow

Educational Profile: -

* **Master of Engineering Computer Science (M.E)** in Sri Krishna College of Engineering and Technology, Coimbatore during 2012 till 2014.
* **Bachelor of Information Technology (B. Tech)** in SNS College of Engineering, Coimbatore during 2008 till 2012.

**Technical Skills:-**

**Cloudera Ecosystems**:Cloudera Hadoop, Spark with Scala, HDFS, Hive, Impala, Zookeeper, Hue, Sqoop, Oozie.

**NoSQL Database**: HBase, Accumulo, Cassandra, Elastic Search

**Databases**:Hive, Impala, Teradata, Hadoop, DB2, MariaDb, Oracle

**Languages**:Core Java, Scala, Python, Spring Boot, JScript’s, Shell Scripts

**Elastic Log & Search**:ELK and Solar Index

**Real-time Ingestion**: Flume, kafka, Custom JAVA Messaging Queue

**Reporting**: R, Qlikview, Kibana, AirFlow, IBM Tivoli

**ETL Tools**: Informatica, Talend, Informatica BDM

**DevOps Tools**: Jenkins, GitHUB, Artifactory, CI/RA, JIRA

**Project Profile:-**

**Project #7**

**Project Name : Power to Predict**

**Company : UOB BANK**

**Tools** : **Spark, Cloudera Hadoop, Java, Scala, Oracle, Shell Scripting, Informatica BDM**

**Role : Big Data Developer**

**Project Description :**

Power to predict to analyse the Bankwide Loyalty Program to extract data from various Fund transaction Sources as Visa, Amex, CUB,..with Merchant Information And from various channels to predict loyalty for individual customers.

**Key Responsibilities:**

* Implemented **Spark Analytical Datamart** using Scala and Java
* Build the **Star schema and SnowFlake schema based OLAP Cubes** accordingly for the reports to be generated for the projects
* **Schema generation and Optimized queries** are generated in Hive on MapReduce, Hive on Spark, Impala, SparkSQL. Readily available for the loaded tables in Impala and Spark
* Co-ordinate with Business Users to get the requirements and to deliver the expected results with the Open Source Platform cost-effectively
* Connect with various source to get Data and Analyse it to Map into the Datamart
* Apply suitable Tropologies to get desired Output from the source Data
* Build the Datamart using Spark with Scala and Java

**Project #6**

**Project Name : Data Ingestion Framework**

**Company : DBS BANK**

**Tools** : **S3, Talend**, **Spark, Hadoop, Hive, Linux, HDFS, ELK,**

**AirFlow, MariaDB, RabbitMQ**

**Role : Big Data Developer**

**Project Description :**

Established the DBS BankWide Enterprise DataIngestion Framework to ingest data from various source to Staging using TALEND Tool for ETL Operations and Big Data Computations. Data launches at S3 bucket using Framework and using Airflow a programtic scheduling and monitoring tool which sense the File arrival on S3 bucket and trigger Talend API to ingest Data to various platforms as HDFS, HBase and SQL Databases.

**Key Responsibilities:**

* Design and Develop **Talend Tool Mappings** (for ETL Processing)
* Completed POC on Talend based on ETL and CI/CD Evolution perspective.
* Designed **End-to-End Airflow Scheduling and Monitoring** Tool Server Setup, Development and **Integrated with Talend**.
* Configured the **Airflow cluster using MariaDB** as a MetaStore and **RabbitMQ** for the Scheduler Messaging on Top of 3 Clusters.
* Ingested Logs flow from Talend, Spark, Hadoop and AirFlow to **ELK**(LogStash, ElasticSearch and Kibana)
* Analysed the Flowlog from Elastic Search to Create a dashboard using **Kibana**
* Integrated Data Ingestion Framework Talend and AirFlow to CI/CD process using **Jenkins, GitHub, JIRA**

**Project Profile:-**

**Project #5**

**Project Name : BIP ETL & BigData Framework**

**Company : DBS BANK**

**Tools** : **Spark, Scala, Hadoop, Hive, Sqoop, Teradata, Informatica,**

**Unix, HDFS, Flume**

**Role : Junior Big Data Architect**

**Project Description :**

BIP ETL & Big Data Framework incorporate ETL operations, API's with various tools, DataModeling Operation, DataWarehousing concepts as Functions to generate batch or real time respective Jobs and Scripts based on Metadata. Framework tools are Hadoop hive and Impala, Sqoop Import and Export, Flume, Teradata, Unix, Java, Hbase, Informatica. This Framework is defined to Deliver Quality value to various projects like DigiBank India/Indonesia, Project Blue, FAIR, BlueSky, Lantern, URs, etc,…

**Key Responsibilities:**

* Build Re-extraction utility with transformation functionality from Hadoop to various destination using **WEB HDFS, JAVA API,REST API , Sqoop Export using Spring Boot**
* Developed **Hadoop Framework** to handle SCD types, Transactional data
* Integrated **HADOOP and TERADATA data exchange Platform**
* Enhanced **Automated Encryption of HIVE** user layer sensitive column information
* Enabled **Realtime Ingestion** on Framework using **Flume**
* Designed Hadoop Stage history Maintenance on **HIVE for Analytical DataModel.**
* POC on **Finding sensitive attributes** based on column level and row level analysis and build predictive model
* Designed Framework to **parse QlikView log** files and provided user layer on top of **IMPALA**
* Fine tune and optimized **HIVE and Impala performance**
* Enabled **Hive on spark** as optional to use by specify project/Job level
* Experienced in **TDCH to export and import** data between Teradata and **HADOOP & Hive**
* Deployed **Hadoop jobs in Cloudera Distribution** Hadoop cluster
* Used **TSQL Procedure, Macro** to generate complex business logic and views based on metadata
* Provided solution to **achieve complex business logics** in simple steps by using Frame Work

**Project #4**

**Project Name : DIGIBANK SOI**

**Company : DBS**

**Tools** : **Spring Java,Hadoop, Hbase, XML, Teradata, Sqoop**

**Unix,HDFS**

**Role : Big Data Developer**

**Project Description :**

As part of this project to pull Customer and Transactional Data from various sources by using Apache FlUME, HDFS and Parsing using DOM Parser. Scope of the project is to enhance and reduce the Realtime message conception and increase performance by using VTD XML Parser. To load the extracted Data’s into the 40+ various tables using HIVE and TERADATA and enable USER Layer & Application Layer on ADM to analyze customer data for generating the Business Logical Reports

**Key Responsibilities:**

* Designed **Data Ingestion Model** to extract Data from **IBM JAVA Message Queue** to BIP
* **Used VTD XML and Custom logic Parser instead of DOM parser to achieve with minimal execution time**
* Injected extracted Data into HDFS & Teradata using **SQOOP**
* Generated HIVE and Teradata User Layer views on ADM.
* **Automated Loading of Transactional / Non-Transactional** table Data into HIVE from Teradata using SQOOP by reducing time latency of data loading and increased performance by handling multiple scenarios

**Project #3**

**Project Name : ACS(Automated Comunication Servilance)**

**Company : DBS**

**Tools** : **Teradata, Informatica,Unix,HDFS,TWS**

**Role : Big Data Developer**

**Project Description :**

To analyse Traders and brokers Chat Message to Prevent Insider Trading, Monitor illegal activities based on Thread words. Extract data from Bloomberge, Yahoo, Cisco, Reuters via Batch and Real Time Mode.

**Key Responsibilities:**

* Developed **Unix shell script** to pull data from sources.
* Extracted Thread words from user uploaded Oracle external DB to BIP by using **INFORMATICA**
* Developed jobs to **load extracted data to HDFS and Teradata**
* Design, Create and Manage **TWS Tivoli Schedule Jobs** to pull data based on daily, hourly, weekly, Yearly, Public Holidays or Defined User Cut-off Dates.
* Provide Re-extraction Layer loaded Data to the Downstream for the further Analyze.

**Project #2**

**Project Name : URs (User Requests)**

**Company : DBS**

**Tools** : **Hadoop, Teradata, Informatica,Unix,HDFS**

**Role : ETL Developer**

**Project Description :**

User Request to Enhance and Enable the additional features and functionality on exsiting project

**Key Responsibilities:**

* As part of IServe Project application provided audio files to user via **HDFS** and **IMPALA** by providing **WEB HDFS API**
* Developed and Implemented Column Addition, Data Type Modification, User layers and application layer by used various Join conditions.
* **Business logic enhancement** on view on **HIVE and TERADATA**
* Extracted **Active User Directory** (AD) results with categorization and loaded to HDFS as part of Cycle Project

**Project #1**

**Project Name : Online Transaction**

**Company : SKAVA**

**Tools** : **JAVA ,JSP,MS Sql2008**

**Role : Java Developer**

**Project Description :**

Online learning framework analyzed for personalized content optimization to fully leverage historical activities. This keeps lot of historical activities of the portal web site. To address this issue and improve the accuracy prioritizing queuing model with updated frequent pattern tree is introduced to improve better result. Project implemented by front end Java-JSP and back end MS SQL 2008*.*

**Key Responsibilities:**

# Reduced duplicates records by automating RI constraint check with java and DB2 SQL.

# Used mathematical algorithms increase performance on calculation

**Publications:-**

# Published Research Work Entitled “**ONLINE CONTENT EXPLOSION FOR CONTENT OPTIMIZATION USING PRIORITIZING QUEUE WITH UPDATED FREQUENT PATTERN TREE**” in International journal of Computer Science trends and Technology Volum2, Issue1, January-March 2014.

# Published Research Work Entitled **“MINIMIZING PACKET LOSS USING CONGESTION CONTROL SCHEME FOR VIDEO STREAMING”** in International Journal of Research In Computer Applications And Robotics, Volume2, Issue2 February 2014.

**PersonalDetails:-**

**SEX & STATUS :** MALE & SINGLE

**DOB :** 01/01/1991

**NATIONALITY :** INDIAN

I hereby assure you that the information furnished above by me is true to the best of my knowledge.