**Bal Jathaul**

**Lead Data Engineer**

**Mobile: (415) 594-6843**
**Location: Kansas City, MO**
**Email:** jathauljay@gmail.com
**LinkedIn:** https://www.linkedin.com/in/bal-jathaul/

### Tech Stack

Microsoft Azure
PySpark
Python
Spark
Scala
Java
ETL
ADLS
Hadoop
AzureDatafactory
Azure Databricks
Hive
SQL
HBase
Impala
Linux

### Key Skills and Knowledge

Database Design and Integration
Data Warehouse
Data Conversion
Data Migration
Data Pipelines development and deployment

**Domain Expertise & Solutions**
Health care Industry

**Databases**
MS SQL Server
NO SQL HBASE

**Tools**
GIT
Automic
Autosys
Jira
IntelliJ
Eclipse

## Professional Summary

- 7 years of hands-on IT experience, worked as Big Data Developer and handling huge data by optimizing pipelines using tools **Hadoop, Scala**, **PySpark**, **Hive** and **Azure DevOps**.
- Work experience as Lead Azure Big Data Developer and Team lead
- Experienced in managing US team and offshore team as a project lead.

## Project Experience

**Molina Healthcare** **06/01/2017-Present**
**Lead Azure Big Data Engineer**

- Designed and implemented **ETL data pipelines** to process structured data by integrating millions of raw records from various data sources like SQL server, using **Azure Databricks**, **Spark API, SQL**, **Scala** and **PySpark.**
- Used **Scala** functional programming using higher order functions and pattern matching.
- Used **First class functions** in which functions were passed as parameters and return them as values.
- Led a project to migrate legacy systems to a modern data lake, utilizing Cats Effect and ZIO for managing side effects and ensuring robust data ingestion processes.
- Ensured data processing workflows comply with HIPAA regulations through rigorous validation using pure functions.
- Processed data via enhanced Spark SQL and Scala codes to load Hive partitioned tables at enterprise layer and provide data to the client, post importing data from **HDFS** using **HiveQL** for **incremental** and **full load** jobs.
- Led performance tuning and optimization initiatives of data pipelines by implementing components like **repartitioning**, **coalesce** functions, **broadcast** joins to enhance job loads.
- **Migrated data pipelines** from on premise **Cloudera** cluster to **Azure** Cloud using various **Azure services**.
- Created workflows in **Azure Databricks** for each pipeline and migrated hive table storage to **Azure data lake storage** using **delta tables and delta live tables**.
- Deployed workflow jobs using **Azure DevOps** and managed the code version control using Azure **GIT**.
- Built data pipelines to load and transform large sets of structured data.
- Imported data from **HDFS** into Hive using HiveQL.
- Involved in creating Hive tables, loading and analyzing data using **HiveQL**.
- Created **Hive partitioned** tables to improve performance.
- Implemented spark code using **Scala** and **PySpark** for RDD transformations and actions in Spark application.
- To improve performance and optimization of the data pipelines, explored different components like storage levels, **Caching** and **checkpointing**, repartitioning and coalesce functions.
- Optimized data caching strategies based on access patterns. Used **persist()** and **unpersist()** judiciously to manage cached data.
- Organizing data to minimize disk seeks. For example, use columnar storage formats (like **Parquet**) which improve I/O efficiency by reading only the necessary columns.
- To improve performance between the joins, implemented **broadcast** and **bucketed** joins.
- Optimized queries to use different types of joins like **map** side and **bucket** joins for efficient query.
- Built reusable UDF's for partitioning and custom cache UDF's.
- Building and deployment of jar files in test and Prod environment in on prem (cloudera).
- Used **Jenkins**, **Azure DevOPS** and **CI/CD** process for deployment.
- Followed agile methodology especially SCRUM software development process through development, QA and UAT.
- Strong knowledge on the task error handling, task editing and debugging of pipelines when failed in prod and other environments.

- Creation, Automation and scheduling of jobs using shell scripting and AutoSys, Automic.
- Creation of workflows for individual pipelines and testing them in dev and uat environments.
- Used Scrum Agile Methodology in my work (Daily Scrum Meeting, User Stories, points, Sprint Backlog, 1on1 meeting).
- Extensively worked on testing and performed code review for various processes build by different
- internal teams within the project.
- Interacting with Business users to get the as-is process requirements.
- Actively participating in test plan presentations, validating the client requirement and identifying the various scenarios for automation development.

- **Education**

  Keller Graduate School of Management, Kansas City, Missouri
  Master's degree in project management (MPM)

  S.D.M. College of Engineering and Technology, Karnataka, India
  Bachelor's degree in computer science.