AKSHAY NIGAM

☎ : +91-9968830525   ✉: nigamakshay11@gmail.com

Seasoned <u>Data Engineer</u> with strong hold on ETL, Data Analytics, Data lake implementation in on-prem and Cloud using Big Data & Spark.

AREAS OF EXPERTISE		
• **AWS**	• **Python, Scala, Unix, Spark**	• **Tableau**

PROFESSIONAL SUMMARY

- Seasoned technology professional with **Five years and Three months** of experience in the space of Business Intelligence and Analytics with strong exposure towards Data Lake pipeline implementation using technologies like big data, Python, Spark and Scala.
- Developed the complex frameworks around features such as ETL, Data Integration, Datalake ingestion, Audit Balance Control checks, Unit Testing, source file data quality checks before ingestion, Change Data Capture, Strong Entitlement Process in data lake, Orchestration, Error logging & notification process around Failures and SLA checks , auto scaling of processing resources in AWS , code validation frameworks.
- Have **4+** years of hands-on experience in implementing solutions around Data Migration, Data Quality, Data Analytics, Data Integration using various ETL Tools & Big data technologies (Hive, Spark, Scala, PySpark) from On-prem to Cloud-AWS.
- Have **1+** years of hands-on experience in implementing end-to-end credit analysis and reporting solutions using Tableau.
- Presently spearheading with **Deloitte Consulting USI** as a **Consultant** at Gurugram.
- Gained good exposure in domains such as Insurance, Life Sciences & Healthcare and Finance while serving the client such as Anthem Life Sciences and Insurance, Trans america, Experian.
- Have expertise in working with various databases like Oracle, Teradata, SQL Server and MySQL handling various types of file systems like XML, JSON, Parquet and flat files.
- Conversant with AWS components (S3, EC2, EMR, RDS, Redshift, Glue, Athena), Python (Numpy, Pandas), Unix, Apache Spark, RDDs, Spark SQL DataFrames, distributed systems and Hadoop cluster.
- Experience working with Agile development methodologies.
- Showcased strong interpersonal and communication skills due to multi-site experience in the U.S. and India.

EXPERIENCE:

Company Name	Job title	Employment Tenure
Deloitte Consulting US India	Consultant	April 2017 – Till Date
Infosys Limited	Systems Engineer	September 2015 – April 2017

EDUCATION:

B.Tech in Electronics and Communications from Guru Gobind Singh Indraprastha University.

CORE COMPENTENCIES:

ETL design solutions	Data Analytics	Data migration
Data Integration	Data Quality	Data Lake implementations
Cloud-AWS	Strong SQL	Orchestration

DOMAIN KNOWLEDGE:

Insurance	HealthCare	Banking and Financials

TECH STACK:

S. No	Area	Technology Stack
1	Cloud	AWS –Storage: S3 AWS-Databases: Athena, DynamoDB, Redshift, AWS-Compute: EC2, EMR
2	Databases - Relational	Oracle, SQL Server, MySQL
3	Databases-NoSQL	AWS – DynamoDB, PostgreSQL
4	Databases-Columnar	AWS Redshift
5	Big Data / Hadoop Ecosystem	HDFS, Sqoop, MapReduce, Hive, HBase, Spark with Python (**PySpark**)
6	Scripting	Unix shell Scripting
7	Reporting & Visualization	Tableau
7	Programming	Python, Scala
8	Version Control	Git, Bitbucket
9	Orchestration (Scheduling)	Apache Airflow
10	CI/CD Integration	Jenkins
11	Others	WinSCP, Excel, Power Point, Visio, Word, JIRA, Confluence, Slack

PROJECT EXPERIENCE:

Project 1: ARK – ARL Integration
Industry: Insurance Service Line
Service Area: Technology
Role: Data Engineer; (6 months)
Tech Stack: Apache Spark, Python, PySpark, AWS – [**S3, Athena, DynamoDB, EC2, EMR, Glue],** PostgreSQL, Orchestration Tool – Apache Airflow

Client – Trans america

Role: Played the role of Data Engineer to develop the AWS cloud-based data lake to ingest the Insurance data coming from legacy ware house. Implemented the End to End Ingestion Framework, **A**udit **B**alance **C**ontrol (ABC) Framework, validation Framework, Unit Testing Framework and Strong Entitlement Process in the Publication layer (PostgreSQL) to cater to the access needs of various business users.

Responsibilities:
1) Development and delivery of generic integration Framework to pull data from Database based sources into the AWS Datalake.
2) **Orchestrate** the data flow of new sources through various layers (landing, raw, staging, curated and Publication) in AWS using Apache Airflow following all sort of dependencies and without hindering the existing functionality.
3) **Performance tuning –**
   a) **Airflow** – SQL Sensors, custom short circuit operators, Pools, etc
   b) **PySpark-** repartitions, broadcast variables.
4) Scheduling of data processing pipelines using Airflow.
5) **Key deliverables** –
   Data lake Integration: Successfully integrated a new Advisor source – ARK to Advisor Data Lake having 75k advisors amounting to more than $20 Billion in asset under management by these advisors.

Project 2: TA Enterprise Data Lake
Industry: Insurance Service Line
Service Area: Technology
Role: Data Engineer; (15 months)
Tech Stack: Apache Spark, Python, PySpark, AWS – [**S3, Athena, DynamoDB, EC2, EMR, Glue],** PostgreSQL, Orchestration Tool – Apache Airflow

Client – Trans america

Role: Played the role of Data Engineer to develop the AWS cloud-based data lake to ingest the Insurance data coming from legacy ware house. Implemented the End to End Ingestion Framework, **A**udit **B**alance **C**ontrol (ABC) Framework, validation Framework, Unit Testing Framework and Strong Entitlement Process in the Publication layer (PostgreSQL) to cater to the access needs of various business users.

Responsibilities:
6) Development and delivery of complex Ingestion Framework requirements.

7) **Design**, Build, Review for **frameworks** – Audit **B**alance **C**ontrol (ABC) framework, Ingestion Framework, Validation framework, Unit Testing framework using **PySpark** and Pytest.

8) Strong **Entitlement Process** implementation in Publication Layer (PostgreSQL) using User access management principles using Roles, access to users at object level, Jenkins etc.

9) **Orchestrate** the data flow through various layers (landing, raw, staging, curated and Publication) in AWS using Apache Airflow following all sort of dependencies.

10) **Performance tuning –**

    c) **Airflow** – bash, custom short circuit operators, Pools, etc

    d) **AWS -** EMR **-**auto scaling, subnet IP ranges

    e) **PySpark-** repartitions, broadcast variables, UDF functionalities, join optimizations, resource optimization and allocation.

11) Estimates system capacity to meet near - and long-term processing requirements

12) **Key deliverables** –

    Data lake: Design for Ingestion, Audit, validation Frameworks, Orchestration Framework design using Airflow.

Project 3: Asset Insights – Scala to PySpark Migration
Industry: Asset Management – Sales & Marketing Service Area
Service Area: Technology
Role: PySpark Developer; (2 months)
Tech Stack: Scala, PySpark, AWS

An Inhouse Accelerator with a business requirement to migrate their Legacy system in Scala to Cloud based data lake (PySpark). The data comprises of Asset Management (Company, Contact, Opportunity, Activities etc.). The goal of this project is to break down each files into Dimension model (XREF, Dimensions, Facts and relation) then perform Match and Merge using PySpark to identify the duplicate company contacts and merge them to single entity.

Role: Pyspark Developer to convert the scala code used in existing systems into PySpark to build the cloud-based data lake to ingest the Asset Management Industry demographic data from various sources.

Responsibilities:
1) Code conversion from Scala spark to Pyspark, adhering to standard guidelines around Data lake Implementations.

2) Creation of complex SQL/HQL in AWS-Athena for structured visualization of Data.

3) Strong Working exposure on different AWS Services (EMR, EC2, Athena, S3) while implementing the data lake in AWS.

4) **Key deliverables** –
    Data lake: Completed the entire codebase conversion from Scala to PySpark in record two month time.

Project 4: Data Processing Engine for a Leading Healthcare firm in US
Industry: Life Sciences and Health Care (LSHC)
Service Area: Technology
Role: Big Data Developer; (12 months)
Tech Stack: Scala, Big Data (HDFS, Hive, Apache Spark)

Role: Big Data Developer; Successfully delivered around 10 complex extracts to production.

Responsibilities:

- Delivered the robust ETL Design solutions in line with quality and Performance metrics established for Data Exchange Extracts in Health care Analytics (HCA) Program.
- Developed a re-usable end-to-end framework using Apache Spark, which processes and transforms data flowing through various layers built on top of Hadoop Distributed systems.
- Implemented automation and scheduling of job pipeline using Control-M
- Implemented Audit Balance Control framework to handle error record based on DQ Rules.

Project 5: Credit scoring and Risk Reporting for Credit Bureau
Industry: Financial Services
Service Area: Technology
Role: Big Data Developer; (12 months)
Tech Stack: Tableau, Big Data (HDFS, Hive), Pentaho (Kettle),

Role: Tableau Developer; Involved in development of end-to-end reporting solution using Tableau.

Responsibilities:

- Developed enterprise reports and dashboards for credit scoring in accordance with the metrics and KPI's provided by the Credit Bureau.
- Conversion of complex data into columnar form, doing ETL operations for data enhancements using Pentaho Data Integration.
- Developed java application to extract and load data from HBase to MySQL DB.
- Creation of Credit Reports for Stakeholders to represent different KPIs and Metrics using variety of chart types in Tableau.
- Worked on Tableau Server for scheduling tableau extract data refresh and also used Shell Utilities for distribution of reports.

REWARDS AND RECOGNITIONS:

Deloitte - Awards	How many times
Applause Award	Once
Spot Award	Thrice

PERSONAL DETAILS:

Date of Birth	7th September
Place of Work/Preferred Location	Delhi / Gurugram, Noida
Marital Status	Single
Hobbies	Playing Football & Cricket; Reading about Financial Market