Arvind Singh Kamlakar

(269) 271 9086 | Pittsburgh, PA | a.kamlakar@gmail.com | https://www.linkedin.com/in/arvindsinghkamlakar/

Data Engineering Professional with **17+ years** of experience Leading/Managing **Data Engineering** and **DWH/BI** Projects. Formulated Project Planning, Resource Loading, Business Requirements Gathering, Solution Architecture, Technical / Functional Design, Process Design, Prototyping, Development, Testing, Training, and Support. Experience integrating and transforming data (**ETL/ELT**) from different sources into **Data Lake, Data Hub, DWH**, and **ODS**. Proven experience in **Data Engineering, Data Warehousing**, and **Data Modeling** Practices.

Technical Skil

OS:	Windows, Linux, Unix
Languages:	Python, Scala, Bash Scripting
Technologies:	Hadoop, Spark, Kafka, Sqoop, Ranger, Atlas, Airflow, Presto
ETL/BI Tools:	dbt, AWS Glue, Databricks, Informatica, IBM Cognos, Tableau
Hadoop Distribution:	Hortonworks, Cloudera, Amazon EMR
RDBMS/NoSQL/MPP:	Snowflake, BigQuery, Hive, HBase, Teradata, Oracle, SQL Server, DB2, MySQL
Query Language:	HQL, T-SQL, PL/SQL, BTEQ
Other Concepts/Tools:	AWS, Azure, GCP, HDFS, S3, YARN, Tez, REST API, CI/CD, Terraform, Docker, Kubernetes, Argo CD, DevOps, GitLab, UCBuild, UCDeploy, Jenkins, Postman, SMILE CDR, FHIR, Agile (Scrum), Tivoli, Control-M, JIRA, Confluence, MS Visio, dbt-expectations, AWS Lambda, Apache Iceberg, Datadog, CloudFormation, Artifactory, Trino, SqlDBM

Certifications

- **Google Cloud Certified - Professional Data Engineer**
- **AWS Certified Data Analytics - Specialty**
- **AWS Certified Solution Architect - Associate**
- **Microsoft Certified - Azure Fundamentals**

Education

Bachelor of Technology in **Electronics and Communication Engineering** in **2005 from Madan Mohan Malviya Engineering College, Gorakhpur,** affiliated to **UP Technical University, Lucknow, India.**

Professional Experience

Staff Data Engineer at Indeed Inc **Pittsburgh, PA (Mar 2022 – Till Date)**

- Designed and Developed **Batch Data Ingestion** pipelines using **Python, Athena, Presto, Docker, Kubernetes, Snowpipe, Terraform,** and **Airflow** for Jobs Postings and Job Seeker's Profile Changes.
- Designed and Developed **Streaming Data Ingestion** pipelines using **Kafka Connect, AWS Lambda, dbt, Terraform, Docker, Kubernetes, CloudFormation, Snowpipe,** and **Airflow** for Job Seeker's Daily Activities.
- Designed and Developed **Transformation** pipelines using **dbt on Athena/Spark/Snowflake, Docker, Kubernetes,** and **Airflow** for Jobs and Job Seeker Subject Areas.
- Developing **Data Quality** solutions using **calogica/dbt-expectations**.
- Building Enterprise **Data Lake** using **Apache Iceberg** with **Parquet** Format.
- Designed and Developed **Data Models** for Snowflake DWH for Jobs and Job Seeker Subject Areas and maintained them on **SqlDBM** for Analytics and Visualization.

- Creating and Maintaining **AWS, Snowflake**, and **Kafka Connect** resources in **Terraform** using official providers from **Terraform registry.**
- Created **Data Products** using **dbt, Python, Docker**, and **Kubernetes** for **Indeed IDW Marketplace**.
- Setup Application Testing and Deployment as **containerized applications** using **Gitlab CI**
- Used **Argo CD** to automate the deployment of application states for **Kafka Connect Server** and **Spark Thrift Server** on **Kubernetes** Cluster.
- Enhancing existing ETL pipelines for Jobs subject areas to reduce their run-time and improve data availability SLO.
- Developed **Datadog** dashboard for **observability and monitoring** of deployed applications and cloud resources.
- Performed Design reviews of the code and data models and provided feedback on areas of improvement.

Senior Big Data Engineer at Highmark Health **Pittsburgh, PA (Oct 2018 – Mar 2022)**

- Leading a team of **5+ Data Engineers** and **Data Analysts** to deliver **Data Engineering** projects using **Agile** Methodology.
- Designed and Developed **Data Ingestion, Transformation,** and **Lineage Frameworks** using **Kafka, Sqoop**, **Spark**, **Hive, Atlas, Python,** and **Shell Scripting** for **Big Data Pipelines, reducing** development **timelines** by **40%**.
- Designed **Data Models** (Snowflake Schema) for **Hive DWH** for Members, Providers, Claims, and Clinical Subject Areas.
- Independently reengineered **HQL** data pipelines to **pySpark** applications using **Data Frames** and **Spark SQL** for large volume claim data processing, reducing runtime by **60%**.
- Collaborated with the **Data Science** team to build **Data Cleansing** pipelines for **ML Model** Training.
- Automated legacy reporting system using **HDFS, Hive, Hbase,** and **Tableau** to provide **Actionable Insights** about **Market Analysis** and **Trends**.
- Ingested files like **JSON, XML, FHIR, HL7, ORC, Avro, Parquet,** and **SAS Datasets** to **HDFS Data Lake** and designed relevant data store formats for **Hive** and **Spark** Data Analytics purposes.
- Developed **ETL Module** endpoint for **REST APIs** using **Javascript** to **POST** data to **SMILE CDR** repository for **FHIR** resources.
- Developed **Audit/Balance/Control** and **Restart Ability** mechanism for **Data Pipelines** and Developed **Operational Dashboard** for SRE Team using **Power BI,** which **reduced Monitoring effort** by **70%**.
- Designed **CI/CD** pipelines using **GitLab**, **UCBuild/UCDeploy**, and **DevOps** Engineers.
- Conducted **SIT**, **UAT,** and **Production Migration** along with **Post Production Support**.

Manager – Data Engineering at Cognizant Technology Solutions Newark, NJ (Oct 2015 - Oct 2018)
Client: Horizon Blue Cross Blue Shield of New Jersey, USA

- Performed **Project Planning, Estimation, Capacity Planning,** and **Resource Loading** for multiple Projects.
- **Managed** a team of **10+ Data Analysts, Data Modelers,** and **Data Engineers** to build a **Hadoop Data Lake** and **Data Warehouse** using **Teradata, HDFS, Hive,** and **Informatica PowerCenter** for the client in an **Onshore/Offshore** setup.
- Conducted **Performance Reviews**, **Hiring,** and **Promotion Evaluations** for Direct Reports.
- Provided **Technical Leadership** to prepare **High-Level Design** documents per **Technical Requirements** provided by Business Analysts.
- Worked with **Data Architects/Modelers** to develop **Data Models** (**Star/Snowflake**) for Members, Providers, Enrollments, Claims, Clinical, and Authorization subject areas.
- Worked with the **Data Governance** team to implement **Data Quality** checks in **ETL/ELT pipelines**.
- Designed **Data Ingestion** and **Aggregation** solutions using **Sqoop, Spark, Hive, Python,** and **Shell Script** to perform large-scale **Data Aggregation** using the **Hortonworks HDP** Platform.
- Designed **ETL** solutions using **Informatica Power Center** for **Change Data Capture** (**CDC), Slowly Changing Dimension (SCD)**, **Late Arriving Dimensions,** and **Data Recycling**.
- Implemented **Push Down Optimization** on **Teradata** to improve **ETL** runtime by **50%** using Teradata utilities such as **MLOAD, FLOAD, FASTEXPORT,** and **TPT**.
- Provided **Weekly Status Updates** to the Client Leadership Team for all projects, **Defect Triaging**, **SIT/UAT Support**, **Production Migration,** and **Post Implementation Support**.

Assistant Consultant at TATA Consultancy Services **Galesburg, MI (Mar 2008 - Sep 2015)**
Client: Eaton Corporation, USA

- Conducted **Requirements Gathering, Effort Estimation, Project Planning,** and **Solution Design** signoffs.
- Migrated **Enterprise Data Warehouse** from **Oracle 10g** to **Oracle Exadata** and **ETL Pipelines** from Legacy ETL tool to **Informatica PowerCenter,** which reduced runtime by **1/4th** and maintenance effort by 2 FTE.
- Designed **ETL** and **BI Solutions** using **Informatica PowerCenter** and **IBM Cognos** for Operational near **Real-Time Reporting**.
- Created **Data Models** for **Extract**, **Staging**, **Transformation,** and **Audit** Layers and developed **ETL Mappings** and **Workflows** using **Informatica PowerCenter**.
- Improved **Algorithms** for Change **Data Capture (CDC)** technique to load **Fact Tables** faster using **Informatica PowerCenter**.
- Implemented solutions for **Slowly Changing Dimensions (SCD)** to load **Dimension tables** using **Informatica PowerCenter**.
- Created **Shell Scripts** for **ETL pre-processing** and **scheduling workflows** using **Control-M**.
- Developed **Cognos Framework Manager** Model (3 Tier Architecture), **Report Studio** reports, and **Transformer Models** (.pyj) for multidimensional Cube.
- Conducted **SIT/UAT** with Analysts/Business Users. Performed **Production Migration** and provided **Post Implementation support**.
  Conducted Training programs on **Informatica PowerCenter** and **IBM Cognos** for new joiners and junior developers.

Software Engineer at Sopra Steria Ltd **Noida, India (Nov 2005 - Feb 2008)**
Client: Boots PLC, UK

- Migrated **Cognos** Impromptu Catalog to **Cognos** Framework Manager.
- Migrated transformer model based on IQDs, CSV, and TXT to **Cognos FM-based** packages.
- Migrated IMRs (Impromptu Report) to **Report Studios** reports with Prompts and Drill-through capabilities.
- Developed and maintained **Transformer Models** based on IQDs, CSV, and text files.
- Performed **Unit Testing** and Conducted **User Acceptance Testing** with Business Users.
- Migrated all Packages, reports, and Cubes to production and provided warranty support.