# Krishna. P

**Data Engineer**

**Professional Summary**

* Over **10+ years** of working experience in the IT Industry as a Data Engineer with the help of big data, Cloud technologies with proficiency in Data Modeling/Data Analysis.
* Extensive use of **cloud computing infrastructure** such as **Amazon Web Services (AWS), GCP,** and **Azure**.
* Adept Software Development Life Cycle **(SDLC)** and various methodologies such as **Agile** and **Waterfall**.
* In-depth knowledge of. **Snowflake Database** and Table Structures.
* Good exposure to **Python programming.**
* Strong experience with **Big Data** **processing** using **Hadoop technologies** **Map Reduce, Apache Spark**, **Hive, and Pig.**
* Good understanding of cloud configuration in **Amazon web services** (**AWS**).
* Good experience in writing Spark applications using **Python** and **Scala**. o Experience working **with SQL, PL/SQL,** and **RDBMS**/**NoSQL** **databases** like **Microsoft SQL Server, Oracle**, **HBase,** and **MongoDB**.
* Created custom transformations and SQL queries in DBT to extract, transform, and load data from various sources into a centralized data warehouse.
* Experience in using the **spark** **application** **master** to monitor the **spark** jobs and capture the logs for the spark jobs.
* Extensive experience in using **ER modeling tools** such as **Erwin** and **ER/Studio, Teradata,** and **Netezza**.
* Good knowledge of **streaming applications** using **Apache Kafka.**
* Experience in designing both time-driven and data-driven automated **workflows** using **Oozie**.
* Hands-on experience with data acquisition into **Hadoop cluster** using **Sqoop**.
* Experience with **ETL** working with **Hive** and **Map-Reduce.**
* Experience in Working **Dimensional Data modeling, Star Schema/Snowflake schema, and Fact & Dimensions Tables.**
* Excellent technical and analytical skills with a clear understanding of design goals of **ER modeling** for **OLTP** and **dimension modeling** for **OLAP.**
* Having experience in developing a **data pipeline** using **Kafka** to store data in **HDFS**.
* Experience in **optimizing Hive queries** by tuning configuration parameters.
* Experienced in performing real-time analytics on HDFS using **HBase.**
* Knowledge of **Data Cleaning, Exploratory** **Data Analysis**, **Data Visualization**, and Data Mining using **SQL/Python.**
* Capable of DAX Expressions, **Power BI Power Pivot, and Power** integrated with Share Point, and in creating dashboards in Power BI and Tableau.
* Good experience in designing jobs and transformations and loading the data sequentially & parallel for initial and incremental loads in **Talend**.
* Experience in **Data transformation, Data mapping** from source to target **database schemas, and Data Cleansing procedures**.
* Real-time experience in **Hadoop/Big Data** related technology experience in **Storage, Querying, Processing, and analysis of data**
* Experience in developing and scheduling **ETL workflows** in **Hadoop** using **Oozie**.
* Proficient in **Data extraction, Data cleaning, Data Loading**, **Statistical** **Data Analysis**, **Data Wrangling, and Predictive** **Modeling** using **Python**.
* Performing extensive **data profiling** and **analysis** for detecting and correcting inaccurate data from the databases and to track **data quality**.
* Experience in designing a component using **UML Design-Use Case, Class, Sequence,** and **Development**, Component diagrams for the requirements.
* Strong experience in using **Excel** and **MS Access** to dump the data and analyze based on business needs.
* Good expertise and knowledge with the **UNIX commands** like changing the permissions of the file to file and group permissions.
* Good communication skills, work ethics, and the ability to work in a team efficiently with good leadership skills.

# Technical Skills

* **Big Data tools:** Hadoop 3.3, HDFS, Hive 3.1.2, Kafka 3.0, Scala, Oozie, HBase2.3, Sqoop 1.4 o **Data Modeling Tools:** Erwin 9.8/9.7, ER/Studio V17,
* **Cloud Services:** Azure DevOps, Azure Synapse, Azure Data Lake, Azure Data Factory and, AWS, Amazon Redshift, Kinesis, GCP, and BigQuery.
* **NoSQL Databases:** HBase and MongoDB o **Project Execution Methodologies:** JAD, Agile, SDLC, Waterfall, and RAD o **Database Tools:** Oracle 12c/11g, Teradata15/14, Netezza, SQL Server, MySQL, Snowflake DB. o **Reporting tools:** SQL Server Reporting Services (SSRS), Tableau, Crystal Reports, Strategy, Business Objects o **ETL Tools:** SSIS, Informaticav10. o **Programming Languages:** SQL, T-SQL, UNIX shells scripting, PL/SQL.
* **Operating Systems:** Microsoft Windows 10/8/7, UNIX

# Work Experience

## Walmart, Illinois Aug 2021 to Present

## Sr. Snowflake Data Engineer

**Responsibilities** o As a **Sr. Data Engineer** I am responsible for all data management-related aspects of a project. o Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.

* Involved in various phases of the **Software Development Life Cycle (SDLC)** as requirement gathering, **data modeling**, analysis, architecture design & development for the project.
* Worked in an **Agile** environment and used rally tool to maintain the user stories and tasks.
* Migrated **data** **warehouses** to **Snowflake** **Data** **warehouse**.
* Responsible for **data governance** rules and standards to maintain the consistency of the business element names in the different data layers.
* Responsible for developing a **data pipeline** with **Amazon AWS** to extract the data from weblogs and store it in **HDFS.**
* Responsible for **Big data** initiatives and engagement including **analysis, brainstorming, POC, and architecture.**
* Extracted files from **MongoDB** through **Sqoop** and placed them in **HDFS** and processed.
* Configured **Spark** streaming to receive real-time data from **Kafka** and store the stream data to **HDFS**.
* Applied **Spark** **streaming** for real-time data transformation.
* Automated data workflows and orchestration using DBT and scheduling tools like Airflow or Cron to ensure regular updates and availability of fresh data.
* Architect and designed the data flow for the collapse of 4 legacy data warehouses into an **AWS** **Data** **Lake.**
* Installed and configured the **Hadoop** **cluster** using **Amazon** **Web Services (AWS)** for **POC** purposes.
* Connected to **AWS Redshift** through **Tableau** to extract live data for **real-time analysis**.
* Used **AWS** **S3** **Buckets** to store the file and injected the files into **Snowflake** **tables** using **Snow** **Pipe** and run deltas using **Data** **pipelines**.
* Optimized data pipelines in DBT for performance and scalability by tuning SQL queries, leveraging indexing strategies, and utilizing caching mechanisms.
* Worked on complex **SNOW SQL** and **Python** Queries in **Snowflake**.
* Configured the above jobs in **Airflow**.
* Created multiple dashboards in **Tableau** for multiple business needs.
* Worked on **MongoDB** for distributed storage and processing.
* Used **MongoDB** to store processed products and commodities data, which can be further down streamed into web applications.
* Experience in moving data between **GCP** and **Azure** using **Azure Data Factory**.
* Implemented data quality checks and validation rules in DBT to ensure accuracy, completeness, and consistency of data.
* Worked on analyzing Hadoop cluster and different Big Data analytic tools including **Pig**, **Hive**, HBase database and Sqoop
* Experience in building and architecting multiple **Data pipelines**, end to end **ETL** and ELT process for Data ingestion and transformation in **GCP** and coordinate tasks among the team.
* Worked on **normalization techniques. Normalized** the data into the **3rd Normal Form (3NF).**
* Involved in **Data profiling** to detect and correct inaccurate data and maintain the **data quality**.
* Responsible for storing processed data in **MongoDB**.
* Responded to and resolved access and performance issues. Used Spark API over **Hadoop** to perform analytics on data in **Hive**
* Build data pipelines in airflow in **GCP** for ETL related jobs using different airflow operators.
* Worked on **SaltStack** automation tools. Helped teams working with batch-processing and tools in Hadoop technology stack **(MapReduce, Yarm, Pig, Hive, HDFS**)
* Collected data using **Spark** Streaming from **AWS S3** bucket in near-real-time and performs necessary Transformations and Aggregations to build the **data model** and persist the data in **HDFS**.
* Responsible for creating **Hive** **tables**, loading data, and writing **Hive** **queries**.
* Worked on **User Defined Functions** in **Hive** to load the data from **HDFS** to run aggregation functions on multiple rows.
* Used cloud shell SDK in **GCP** to configure the services **Data Proc**, Storage, **BigQuery**
* Created various types of **data visualizations** using **Python** and **Tableau**.
* Written and executed **unit, system, integration,** and **UAT scripts** in a data warehouse project.
* Coordinated with **SCRUM** Master in delivering agreed user stories on time for every sprint.

**Environment**: Hadoop 3.0, Agile/Scrum, Spark, AWS, Snowflake, Python 2.7, Hive, Tableau, HDFS, Airflow, Mongo DB, Kafka 2.8

## Kaiser, Oakland, California June 2017 to July 2021

## Sr. Data Engineer

**Responsibilities**

* As a Data Engineer Reviewed Source to transformation mapping (STM) documents to understand the functionality and requirements and was involved in gathering business requirements.
* Involved in various phases of development analyzed and developed the system going through **Agile Scrum methodology**.
* Acted as a lead resource and built the entire **Hadoop platform** from scratch.
* Worked closely with the **DevOps team** to understand, design, and develop end-to-end flow requirements by utilizing **Oozie workflow** to do **Hadoop jobs.**
* Installed and configured **Hadoop MapReduce, HDFS**, and developed multiple **MapReduce** **jobs** **Scala** for data cleaning and preprocessing.
* Worked on **Data governance, data quality, and data lineage** establishment processes. o Responsible for ingesting data from various source systems **(MySQL, Flat files,** **Big Data**) into **Azure (Blob** **Storage)** using a framework model.
* Involved in building **data pipelines** using **Azure Data Factory, Azure data bricks** and loading data to **Azure Data** **Lake and** Controlling and granting database access.
* Designed and implemented a fully operational production-grade large-scale data solution on **Snowflake Data** **Warehouse**.
* Implemented version control and deployment strategies using Git and CI/CD tools to manage DBT codebase and ensure smooth deployment of data models and pipelines.
* Developed **MapReduce** programs to parse the raw data, populate staging tables, and store the refined data in **partitioned tables** in the **EDW.**
* Provided End to end business intelligence solutions by using Microsoft technologies using **Azure Data** **Factory**, and **Azure Synapse**.
* Configured **Spark streaming** to get ongoing information from **Kafka** and store the stream information to **HDFS**.
* Worked with analysts to determine and understand **business requirements.**
* Involved in composing shell scripts to dump shared information from **MySQL** servers to **HDFS.**
* Worked on Implementing and optimizing **Hadoop/MapReduce algorithms** for **Big Data analytics.**
* Developed data integration strategies for data flow between disparate source systems and **Big Data** enabled.
* Created **data integration** and technical solutions for **Azure** **Data** **Lake** for providing analytics and reports for improving marketing strategies.
* Involved in porting the existing on-premise Hive code migration to GCP (Google Cloud Platform) BigQuery.
* Extracted data using **Sqoop** Import query from multiple databases and ingested into **Hive tables.** o Designed and built a **Data Discovery Platform** for a large system integrator using **Azure HDInsight** components.
* Involved in writing **MapReduce code using Python** to get rid of certain security issues in the data.
* Used **Azure Data Factory** as an orchestration tool for integrating data from upstream to downstream systems.
* Integrated **Apache Kafka** for data ingestion.
* Responsible for setting up **HBase** and storing data in **HBase.**
* Modified Database tables and used **HBase Queries** to insert and fetch data from tables.
* Developed **Python utility** to validate **HDFS** **tables** with source tables.
* Used **Spark** as an **ETL tool** to remove Duplicates, Joins, and aggregate the input data before storing in a Blob.
* Encoded and decoded JSON objects using **PySpark** to create and modify the data frames in **Apache Spark.**
* Evaluated and improved application performance with **Spark**.
* Designed and developed automation test scripts using **Python**.
* Well-versed in data engineering, utilizing **EMR, PySpark, Redshift, Glue**, and serverless technologies such as **Lambda** and **step functions**. Also experienced in containerization using **ECS** with **Fargate**.
* Used **Azure Data Bricks** to do **ETL** **(Workspace, Hive, RDD, MapReduce** so on**).**
* Worked with **Spark** for improving performance and optimization of the existing algorithms in **Hadoop** using **Spark Context**, **Data Frames and Pair RDD**'s.
* Used **Jenkins** like **CI/CD tools** to make daily builds and deploys.
* Implemented Custom **Azure Data Factory Pipeline** Activities and SCOPE scripts.
* Involved in converting **Hive/SQL queries** into **Spark transformations** using **Spark RDDs, Python**.
* Proficient in **Python** with strong experience, including working extensively with **AWS** and developing user interfaces using **JSP** and **Angular**.
* Developed **Databricks Python notebooks** to Join, filter, pre-aggregate, and process the files stored in Azure data lake storage.
* Designed and implemented data lineage and documentation using DBT to track the flow of data and enhance data governance and data cataloging.
* Proficient in testing automation to ensure software quality.
* Scheduled Automatic refresh and scheduling refresh in **Power BI service.**
* Involved in performance tuning and monitoring of **T-SQL blocks.**
* Coordinated with **SCRUM** Master in delivering agreed user stories on time for every sprint.

**Environment:** Hadoop3.3, Agile, Azure, Talend, Python3.10, Power BI, MySQL, HBase2.3, Kafka3.0, Snowflake Data Warehouse, Hive3.1.2, Pig0.17, T-SQL, Oozie, MapReduce, Spark3.2, Azure DevOps, HDFS, OLTP, CI/CD, Jenkins, Sqoop, Azure Synapse, Big Data.

## Biztegy Analytics, Hyderabad, India Oct 2014 to May 2017 Data

Role: Java Full Stack Developer

Responsibilities:

* Involved in Scrum (daily standups) meetings, Sprint/retrospective planning and estimation of the tasks for the user stories and refine the tasks for next sprint.
* Involved in various **SDLC** phases like **Requirement** gathering, **Design**, **Analysis** and **Code development**.
* Used **Java 8** Features Like **lambda expressions** for communicating between business layers and database, **stream API** for Bulk Data Operations on **Collections** using both streams and parallel streams.
* Excellent experience of using **NPM code Packages** and managing **JavaScript** projects.
* Have extensive knowledge on **JavaScript/TypeScript** dynamic programming languages. (Async await, HTTP requests etc.)
* Built applications using **React JS** with **ES6** which allowed us to render pages on both the client and server using the same codebase, developed web and mobile applications using ES6 features and frameworks such as React, React Native and Redux.
* Pleasant Experience in **React.js** for creating interactive UI's using One-way data flow, **Virtual DOM**, **JSX**, **React Native** concepts.
* Strong in using the **Node.js**, and Require.js and MVC Frameworks in the development of the web application.
* Experience in User Experience (UX) design and development, User Interactive (UI) Web Pages and visually appealing User Interfaces using **JavaScript,** **Vue.js**, **Typescript**, Node.js and **React.js**.
* Developed UI pages with **HTML5**, **CSS3**, **JavaScript**, **Bootstrap** for creating Responsive web pages.
* Extensive experience in developing **Microservices** using **Spring frameworks** and followed domain driven design.
* Experience in using build/deploy tools such as **Jenkins, Docker and OpenShift** for Continuous Integration & Deployment for **Microservices**.
* Deployed **Spring Boot** based **Micro services Docker** containers Using **AWS EC2 containers** services and using **AWS admin console**
* Implementing or exposing the **Microservices** based on **RESTful API** utilizing **Spring Boot** with **Spring MVC**.
* Implemented **Restful Web Services** in **Spring Boot** for consuming, producing JSON/XML Data from End points.
* Good knowledge of **Web Services** with protocols **SOAP** and **REST**.
* Developed the Spring Features like **Spring MVC**, **Spring Boot**, **Spring Batch**, **Spring Security**, and **Spring Integration**.
* Secured **Rest API’s** by implementing **OAuth2** tokens-based authorization schema by **Spring Security**.
* Experience in working with **AWS**, **EC2,** and **S3, Cloud watch platform**. Created multiple VPC, Subnets in AWS as per requirements.
* Understanding the Customer requirements and master franchises future developments to their markets, develop exact to their requirements.
* Managed database clusters and ensured their high availability, implementing replication and sharding strategies for seamless data distribution.
* Developed custom data models and optimized queries, leading to a 25% reduction in overall database storage consumption.
* Collaborated with cross-functional teams to integrate NoSQL databases with existing systems, enabling smooth migration from legacy databases.
* Implemented data backup and disaster recovery procedures, ensuring data integrity and business continuity in case of failures.
* Conducted performance tuning and profiling, identifying bottlenecks and implementing optimizations that improved system response times by 20%.
* Contributed to the open-source NoSQL community by creating and sharing database-related tools and utilities on GitHub.
* Mentored junior database engineers, providing technical guidance and fostering their understanding of NoSQL concepts and best practices.
* Created/managed **Mongo DB/NoSQL** collections. Developed Server-side automation using **Node.js** scripting and connecting different types of **SQL** and **NoSQL databases**.
* Making documents and training manual for users in multi languages as required for project.
* Used IntelliJ for development and debugging **Spring Boot** applications and Used **GIT** for version control.
* Developed **CI/CD** pipeline and build tasks to run the unit tests on **Jenkins** and maintained separate build in **Maven** for the module for easy integration to the portal build and configure with **Jenkins**.
* Used **GIT** as version management, Jira for defect management system.
* **Unit testing** and **E2E testing** frameworks are used.
* Supports different markets for alerts and bugs, Analyze and understand fix them.

**Environment:** Java8, TypeScript/JavaScript, React JS, Node JS, Spring boot, Spring security, Microservices, SQL, PL/SQL, NoSQL, webservices, Mongo DB, GIT, Jira, VS code, AWS, S3, Cloud watch, Docker, Jenkins CI/CD

## BilvaSoft, Hyderabad, India Sep 2013 to Oct 2014 Data Analyst/Data Modeler

**Responsibilities**

* Worked as a **Data Analyst / Data Modeler and** I was responsible for all data-related aspects of a project.
* Created **Source Target Mappings (STM)** for the required tables by understanding the **business** **requirements** for the reports.
* Developed a **Conceptual Model** and **Logical Model** using **ER/Studio** based on **requirements analysis**.
* Generated parameterized queries for generating tabular reports using **global variables, expressions, functions,** and **stored procedures** using **SSRS.**
* Created mappings using pushdown optimization to achieve good performance in loading data into **Netezza.**
* Developed **Data mapping, Transformation**, and **Cleansing** rules for Data Management involving **OLTP** and **OLAP**.
* Designed both **3NF data models** for **OLTP** systems and dimensional data models using **star** and **snowflake.**

**Schemas**.

* Worked on **Informatica Utilities** Source Analyzer, warehouse Designer, Mapping Designer, Mapplet Designer, and Transformation Developer.
* Created a **Data Mapping** document after each assignment and wrote the transformation rules for each field as applicable.
* Involved in developing **Unix Shell Scripts** for automation of **ETL process.**
* Worked on data transformations and **data quality** rules.
* Worked on **Data Mining** and **data validation** to ensure the accuracy of the data between the warehouse and source systems.
* Worked on **Normalization** and **De-Normalization** techniques for **OLAP systems.**
* Developed and presented **Business Intelligence reports** and product demos to the team using **SSRS (SQL Server Reporting Services).**
* Designed and Developed **PL/SQL procedures, functions,** and **packages** to create Summary tables.
* Worked on visualizing the reports using **Tableau**.
* Familiarity with **SAS** is a desirable skill, along with knowledge of DevOps practices involving **Jenkins, Bitbucket**, and infrastructure provisioning tools like **Terraform, UCD**, and **CloudFormation**.
* Worked on Performance Tuning of the database which includes indexes, optimizing **SQL Statements.**
* Executed change management processes surrounding new releases of **SAS functionality**.
* Prepared **complex T-SQL queries, views**, and **stored procedures** to load data into staging area.
* Participated in **data collection, data cleaning, data mining,** developing models and visualizations.
* Wrote **complex SQL queries** for **validating** the data against different kinds of reports generated by **Business Objects.**
* Used **Excel** with **VBA** scripting to maintain existing and develop new reports as required by the business.

**Environment:** ER/Studio, Netezza, SAS, SSRS, SQL, Tableau, UNIX, OLAP, OLTP, VBA, MS Excel.