A blue and white logo

Description automatically generated**Premchand Jayachandran**

Senior Data Engineer

**Overall Summary:**

* 8+ years of overall experience as a Data Engineer, Data Analyst, and ETL Developer, involved in designing, deploying, and maintaining **data models** in production systems.
* Proficient in various data engineering steps like collecting data, cleansing data, transforming data, along with automating these steps and monitoring the automated jobs.
* Good experience in programming languages like **Python, SQL, R, Scala**, and **Spark.**
* Expertise in AWS services such as **Redshift, Glue, Kinesis, EMR, CloudWatch,** and **Athena**.
* Strong understanding in different database technologies like **MySQL, Postgres, MongoDB, AWS RDS,** and **AWS DynamoDB**.
* Hands-on experience in various Data Warehousing technologies and Data Lake technologies like **AWS S3, AWS Redshift, AWS Lake Formation, Oracle,** and **Snowflake**.
* Accomplished complex **HiveQL** queries for required data extraction from **Hive** tables.
* Extensive experience in Hadoop led development of enterprise level solutions utilizing Hadoop components such as **Flume, Pig, HBase,** and **YARN**.
* Experience in designing and creating RDBMS Tables, Views, User Created Data Types, Indexes, Stored Procedures, Cursors, Triggers and Transactions.
* Experience in using various packages in **Python** like **Pandas, Numpy, Scikit-learn, Seaborn, Matplotlib, Plotly, SQLAlchemy, PySpark, PyTest, spaCy, NLTK, TensorFlow, Keras, and PyTorch**.
* Experience in Big Data Processing using Hadoop and other tools such as **MapReduce, Sqoop, and Spark**.
* Highly skilled in developing data visualization using tools such as **Tableau** and **Power BI**.
* Implemented data pipelines using **Python** scripts for transformations and **Apache Airflow** DAGs to orchestrate the jobs.
* Proven expertise in designing, developing, and deploying large-scale data warehousing and analytics solutions using **Snowflake**.
* Designed and developed ETL solutions using **Informatica PowerCenter and PowerExchange**.
* Worked with DevOps teams to integrate **Docker**-based solutions into CI/CD pipelines.
* Extensive experience in **designing**, **building**, and **maintaining big data architectures** that are scalable, secure, and cost-effective, using **Spark** and **Kafka.**
* Proven ability to work with **cross-functional teams**, including data scientists, data engineers, software developers, and business stakeholders, to understand their **requirements** and **design architectures** that meet their needs.
* Used **Jira** software to manage **Agile** projects, including backlog grooming, sprint planning, and retrospective.
* Proven experience in designing and executing successful **migration** strategies for **on-premise** data and applications to **AWS cloud**.
* Experience in **designing** data models for **AWS** and knowledge of cloud-specific **data modeling** considerations such as scalability, elasticity, and cost optimization.
* Skilled in building **data models** across multiple platforms such as **SAP, AWS, ER/Studio,** and **ERwin.**
* Expertise in designing and developing **data marts** for business intelligence (BI) and Reporting applications.
* Proficient in using **Spark**'s core APIs, including **RDD**, **DataFrame**, and **Dataset**, and in leveraging Spark's distributed computing capabilities to process and analyze data in parallel.
* Experience in building **Spark** applications using programming languages such as **Python** and **Scala**, and in using Spark's built-in libraries such as **MLlib** and **GraphX** to perform machine learning and graph processing tasks.
* Expertise in using **Databricks** notebooks and its APIs to build and manage end-to-end data processing pipelines that leverage **Apache Spark**'s distributed computing capabilities.
* Extensive experience in data modeling using **SAP BW** extended star schema and entity relationship model. Involved actively in enterprise data warehousing (EDW) strategies.
* Expertise in **ABAP** programming for writing the User exits using CMOD and in writing the start routine, end routine, expert routine and infoobject routine.
* Experienced in creating **functional documentation**, including **functional specifications**, **requirements documents**, **user manuals**, and **training materials** for software and data engineering projects.
* Skilled in **collaborating** with stakeholders from **different functional areas**, including business analysts, developers, testers, and end-users, to elicit and **document functional requirements**.
* Experienced in **managing dependencies** and r**esolving conflicts** between **different teams**, and ability to **anticipate and mitigate risks** that may impact project delivery.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, R, Spark, Scala |
| **Databases** | MySQL, MongoDB, Postgres, Amazon RDS, Amazon DynamoDB |
| **Big Data Technologies** | Hadoop, Hive, Sqoop, Spark, Kafka |
| **ETL Tools** | Informatica, Amazon Redshift |
| **Data Warehouses and Data Lakes** | Amazon S3, Amazon Lake Formation, Databricks, Snowflake, SAP BW |
| **Cloud Technologies** | AWS, GCP, Azure |
| **Machine Learning libraries and algorithms** | Scikit-learn, Numpy, Pandas, TensorFlow, Keras, PyTorch, Seaborn, Matplotlib, Plotly, Linear and Logistic Regression, SVM, KNN, Random Forest, Decision Trees, CNN, Transformers |
| **Visualization Tools** | Tableau, PowerBI, Quicksight, SAP BusinessObjects, Lumira |
| **Orchestration** | Airflow, Step Functions, Cloud Composer, Data Factory |
| **Containers** | Docker |

**Certifications:**

* [AWS Certified Solutions Architect - Associate](https://www.credly.com/badges/8d101f9b-29bd-412a-8a05-d4b6bf7a34c6)
* [AWS Certified Cloud Practitioner](https://www.credly.com/earner/earned/share/7ca9dfcb-8e4c-4041-a47f-132b57fdcbe5)
* [Snowflake SnowPro Core](https://achieve.snowflake.com/ee3e5894-c1ce-4e10-870a-e4499995df27#gs.2oith8)
* [Academy Accreditation - Databricks Lakehouse Fundamentals](https://credentials.databricks.com/6ceb8f08-a816-44d8-9053-f93533774e96#gs.z243eb)

**Fiserv, Alpharetta,GA Dec 2021 - Present**

**Role: Senior Data Engineer**

**Responsibilities:**

* Designed and implemented data pipelines using **AWS Glue** and **Apache Airflow** to extract, transform, and load data into data warehouses.
* Designed and built a robust and scalable **data platform** on AWS, incorporating **Amazon S3** for data storage, **AWS Glue** for **ETL** processes, and **Amazon Redshift** for data warehousing, resulting in a highly efficient and accessible data environment.
* Utilized **AWS'**s powerful serverless capabilities, such as **Lambda** functions and Step Functions, to automate data workflows and enhance the efficiency of the data platform.
* Integrated **Glue ETL** jobs with other AWS services such as **Lambda**, **SNS**, and **SQS** to build end-to-end data processing solutions.
* Designed and implemented **PySpark** pipelines to extract, transform, and load data from various sources, including **Hadoop HDFS, Amazon S3,** and relational databases, ensuring data consistency and quality.
* Developed complex **PySpark** scripts to perform data cleansing, transformation, and feature engineering tasks, resulting in improved data accuracy and suitability for machine learning models.
* Designed and implemented a scalable data warehousing solution using **Amazon Redshift** to store and analyze large volumes of data.
* Developed **ETL** processes using **AWS Redshift** and **PySpark** to perform data transformations and manipulations.
* Developed **Docker** images and containers for **Apache Spark** clusters to perform large-scale data processing.
* Designed and implemented data visualizations using **Tableau** to communicate data insights to stakeholders.
* Maintained data infrastructure using AWS technologies like **EC2, S3**, and **RDS**.
* Developed custom **Airflow** operators and hooks to connect to various data sources and perform data transformations.
* Developed **AWS Lambda** functions and **Amazon API Gateway** for serverless computing and RESTful API development.
* Implemented **AWS Elastic Beanstalk** for easy deployment and management of web applications.
* Utilized **AWS CodeDeploy** for automated application deployment to **EC2** instances or on-premises servers.
* Designed and implemented **Elasticsearch** clusters to store, index, and search large volumes of structured and unstructured data.
* Implemented **Kafka** connectors to stream data from various sources into **Kafka** topics, and from **Kafka** topics into **S3.**
* Used **AWS Glue** in conjunction with other AWS services such as **Redshift, Athena,** and QuickSight to enable ad-hoc querying and data visualization for business analysts.
* Worked on the data migration project, beginning with a comprehensive data mapping and schema design to ensure a seamless transition from **AWS Redshift to Databricks.**
* Conducted a thorough analysis of source data in **AWS Redshift,** identifying data structures and quality issues.
* Implemented data cleansing and enrichment techniques, addressing data quality issues, and enhancing data quality in the new environment.
* Designed ETL processes for data extraction, transformation, and loading into **Databricks**.
* Created custom data visualizations and reports using **Python** packages such as **Matplotlib, Seaborn,** and **Plotly.**
* Developed and deployed machine learning models using Python packages such as **TensorFlow, Keras,** and **PyTorch.**
* Implemented **Snowflake** features like **Snowpipe** for continuous data ingestion and Materialized Views for improved query performance.
* Successfully integrated external APIs with our data platform using AWS API Gateway, AWS Lambda, and Snowflake, enabling real-time data updates and streamlining data workflows.
* Utilized **Jira** to manage project tasks and collaborate with **cross-functional** teams to ensure successful project completion.
* Implemented **ETL** processes **Sqoop**, to extract, transform, and load data into data lakes.
* Orchestrated a comprehensive data processing **POC** leveraging **Azure** services, including **Azure Data Factory** for **ETL** workflows, **Azure Databricks** for scalable data processing, and **Azure Synapse** for data warehousing.
* Integrated machine learning models developed in **Azure Machine Learning Studio** into the POC, demonstrating predictive analytics capabilities for informed decision-making.
* Demonstrated the scalability of the solution by efficiently handling large volumes of data, thanks to the elasticity of **Azure Synapse Analytics** and the distributed processing capabilities of **Azure Databricks**.
* Utilized **Azure EventGrid** to establish an event-driven architecture, enabling seamless communication between microservices and triggering actions based on data events.
* Implemented a robust monitoring and diagnostics framework using **Azure Monitor**, ensuring real-time visibility into the health and performance of the entire solution.
* Built and deployed deep learning models, such as **convolutional neural networks (CNNs)** and **recurrent neural networks (RNNs)**, to extract features from complex data types such as images, text, and audio.
* Utilized **natural language processing (NLP)** tools, including **spaCy** and **NLTK**, to extract insights from unstructured data.
* Collaborated with data scientists to integrate **PySpark** with machine learning frameworks like scikit-learn and TensorFlow, allowing for seamless model deployment and execution at scale.
* Demonstrated end-to-end AI/ML integration with **Vertex AI, BigQuery,** and **Cloud Storage** in a successful **Proof of Concept (POC**).
* Proficiently trained and deployed machine learning models using **Vertex AI**, showcasing **GCP**'s scalability and efficiency in AI services.
* Utilized **BigQuery** for data analytics, revealing insights through SQL-based queries within the AI/ML workflows.
* Efficiently managed and scaled data using **Cloud Storage**, seamlessly integrated within the broader AI/ML architecture on **Google Cloud Platform**.
* Worked closely with data scientists and analysts to understand their data requirements and provide access to the data they needed.
* Trained and mentored junior data engineers on best practices for data engineering and AWS technologies.

**Environment:** AWS Glue, Apache Airflow, AWS Redshift, Apache Spark, Docker, AWS IAM, SQL, Python, AWS CloudWatch, AWS CloudTrail, AWS EC2, AWS S3, AWS RDS, Google Cloud Storage, Big Query, Vertex AI, Tableau, Snowflake, AWS Lambda, AWS API Gateway, AWS Elastic Beanstalk, AWS CodeDeploy, Elastisearch, Azure Databricks, Azure Data Factory, Azure Synapse Analytics, Azure EventGrid, Azure Monitor, Apache Sqoop, Jira.

**Southwest, Dallas, TX Jan 2021 - Nov 2021**

**Role: Senior Data Engineer**

**Responsibilities:**

* Developed a **data platform** architecture on **AWS**, incorporating data lakes, data warehouses, and data streaming capabilities to meet the organization's diverse data needs.
* Design, develop, and maintain highly scalable and efficient data pipelines using AWS services such as **S3, EC2, EMR,** and **Redshift.**
* Configured and optimized **AWS** storage solutions such as **Amazon S3** and **Amazon EBS** for big data workloads.
* Developed **AWS Kinesis Data Analytics** jobs for real-time data processing and analytics.
* Implemented **PySpark** streaming applications for real-time data processing, enabling immediate insights and actions on streaming using **Apache Kafka**.
* Worked with **Amazon DynamoDB** for NoSQL databases.
* Implemented **Amazon Elastic Kubernetes Service (EKS)** for container orchestration and management.
* Developed **AWS CodePipeline** pipelines for CI/CD and infrastructure as code (IaC) for AWS resources.
* Develop and maintain **ETL** jobs using **Python** for data processing and integration.
* Developed and optimized **SQL** code to extract, transform, and load (ETL) large volumes of data from various sources into the data warehouse.
* Designed and implemented complex data pipelines using **SQL** procedures, functions, and packages for efficient data processing and integration.
* Proficient **in PySpark SQL**, enabling SQL-like querying of distributed datasets and simplifying data exploration for business analysts.
* Design and implement data architectures using **Apache Spark** for data processing and analytics.
* Develop and maintain data models for data warehousing using **AWS Glue**.
* Developed custom Py**Spark** functions (UDFs) to address specific data transformation requirements, enhancing the flexibility and adaptability of data processing pipelines.
* Implemented **PySpark** on cloud-based platforms like **AWS EMR** (Elastic MapReduce) optimizing resource allocation and reducing infrastructure costs.
* Continuously optimized the AWS data platform by fine-tuning ETL processes, automating data quality checks, and implementing data governance practices.
* Utilized **AWS CloudWatch** to monitor and log data pipeline performance metrics for analysis and optimization.
* Utilized **AWS X-Ray** for distributed tracing of data pipeline components to identify performance bottlenecks and areas for improvement.
* Utilized **AWS Data Pipeline** for the management of data pipeline workflows and optimized pipeline performance using pipeline activity logs.
* Designed and implemented data marts using **Snowflake**, delivering real-time insights to business units.
* Successfully leveraged **Snowpipe** for automatic data ingestion, ensuring the timely availability of data for reporting and analytics.
* Performed data transformations and maintained data quality within **Snowflake**, ensuring accurate and reliable data in data marts.
* Developed **Snowflake** stored procedures, views, and user-defined functions to perform data transformations and manipulations.
* Designed and implemented **Databricks** clusters and workspaces to enable scalable and collaborative data processing and analytics.
* Created complex **Databricks** notebooks using **Python, Spark,** and **Scala** to manipulate and analyze large datasets, and implemented best practices for performance optimization such as caching, partitioning, and bucketing.
* Utilized **Databricks** to create **ETL** pipelines that integrated data from various APIs into our data lake, ensuring data transformation, validation, and storage according to business requirements.
* Implemented a **Proof-of-Concept (POC)** data pipeline using **Apache Airflow** for orchestrating data transfer from **Hive to MySQL**.
* Developed Python-based **Airflow DAGs** to demonstrate the feasibility of cross-database ETL processes between **Hive and MySQL**.
* Utilized **GitHub** for collaborative development, establishing a centralized repository to manage and store **Airflow DAG** code.
* Orchestrated the **Apache Airflow** job using **Google Cloud Composer**, showcasing seamless integration with **Google Cloud Platform** for efficient workflow management.
* Developed a data pipeline to extract, transform, and load (ETL) data from multiple sources using Python packages such as **Pandas, NumPy,** and **Scikit-Learn**.
* Implemented **PySpark** machine learning workflows, leveraging libraries such as **MLlib** and **Spark ML** to build predictive models for customer segmentation, churn prediction, and recommendation systems.
* Designed and implemented a data model to support efficient data analysis and reporting using Python packages such as **SQLAlchemy** and **Django ORM**.
* Managed **Docker** containers using Kubernetes for container orchestration and scaling.
* Implemented **Apache Airflow** DAGs to automate **ETL** processes, orchestrate workflows, and schedule jobs.

**Environment:** AWS S3, AWS EC2, AWS EMR, AWS Redshift, AWS EBS, AWS Kinesis Data Analytics, AWS DynamoDB, Amazon EKS, AWS CodePipeline, Python, SQL, Apache Spark, Databricks, AWS Glue, AWS CloudWatch, AWS X-Ray, AWS Data Pipeline, Docker, Google Cloud Composer, Apache Airflow, Hive, Apache Sqoop, MongoDB, MySQL, PostgreSQL, PL/SQL, Snowpipe, Power BI.

**Roche, Santa Clara, CA Sep 2018 - Dec 2020**

**Role: Data Engineer**

**Responsibilities:**

* Design and build data pipelines using **AWS Glue** and **PySpark** to process and store large volumes of data.
* Develop and maintain data warehouses using **AWS Redshift** and **SQL** for efficient data processing and storage.
* Implemented **AWS KMS (Key Management Service)** for encryption of data at rest in AWS storage services such as **Amazon S3** and **Amazon EBS**.
* Utilized **AWS S3** bucket policies and access control lists (ACLs) to control access to data stored in Amazon S3.
* Developed custom Lambda functions using **Python** to handle complex data structures and implement business logic for data processing and validation.
* Utilized **AWS Lambda** and **Amazon S3** for serverless computing and storage of data for efficient and cost-effective processing.
* Design and develop data solutions using **Hive** and **Sqoop** to process and store large volumes of data.
* Create **ETL** processes to migrate data from various sources to the data warehouse.
* Created and maintained data models within **Snowflake**, including the development of schemas, tables, views, and stored procedures.
* Implemented efficient **star** and **snowflake schemas** for data warehouse optimization.
* Worked with **AWS Glue Data Catalog** to organize and manage metadata about data sources for efficient data discovery and data governance.
* Automated **ETL** jobs using **AWS Glue** triggers and **AWS Lambda** functions to enable event-driven data processing and reduce manual intervention.
* Designed **PySpark** workflows for **ETL** (Extract, Transform, Load) processes, optimizing data ingestion and transformation pipelines, and reducing data processing time by 30%.
* Utilized **AWS Redshift** for high-performance data warehousing and optimized Redshift clusters for faster query processing and improved query performance.
* Worked with business stakeholders to design and implement **IAM** policies that align with regulatory compliance requirements such as HIPAA, GDPR, and PCI DSS.
* Keep up to date with new technologies and best practices in data engineering and cloud computing.
* Collaborate with cross-functional teams to understand data needs and provide data resources.
* Gained knowledge of data modeling and database design principles.

**Environment:** AWS Glue, PySpark, AWS Redshift, SQL, AWS KMS, AWS Lambda, AWS S3, AWS EBS, Hive, Apache Sqoop, AWS Glue Catalog, AWS IAM.

**Nestle, Bangalore, India Jun 2016 - Aug 2018**

**Role: ETL Engineer**

**Responsibilities:**

* Developed and maintained **ETL** pipelines using **Sqoop** and **Kafka** to move data between on-premises and cloud-based data storage solutions.
* Designed, developed, and maintained **ETL** processes using **Informatica PowerCenter** to extract, transform, and load data from various source systems into data warehouses.
* Created dashboards and visualizations using **Power BI** and **Excel** to communicate insights to non-technical stakeholders.
* Utilized **Hadoop** and **PySpark** to process and analyze large volumes of data to inform business decisions.
* Developed and maintained data models using entity-relationship and dimensional modeling techniques.
* Developed and implemented data quality checks to ensure accuracy and consistency of data.
* Conducted statistical analysis on large datasets using **Python** and **SQL**.
* Created dashboards and visualizations using **Tableau**.
* Worked closely with data analysts and business stakeholders to understand their reporting needs and translate them into effective data visualizations using **Tableau**.

**Environment:** Informatica PowerCenter, Apache Kafka, Apache Sqoop, Hadoop, Spark, Power BI, Tableau, Python, SQL, Excel.

**TVS, Chennai, India Jan 2015 - May 2016**

**Role: SAP Developer**

**Responsibilities:**

* Collaborated with business stakeholders to gather requirements and design **SAP BW** solutions to meet business needs.
* Developed and maintained **ETL** processes to extract data from various source systems into **SAP BW**.
* Implemented data quality processes using **Informatica IDQ** to ensure data accuracy and consistency.
* Developed and maintained master data management (MDM) solutions using **Informatica MDM**.
* Created and maintained data models in **SAP BW**, including **InfoObjects**, **InfoProviders**, and **InfoCubes**.
* Developed and maintained **SAP BW** reports using **SAP BusinessObjects** and **Lumira**.
* Optimized **SAP BW** performance using **aggregates**, **partitioning**, and **indexing**.
* Maintained **ETL** process documentation and provided technical support to end-users.
* Configured **Jira** workflows, fields, and screens to match team-specific processes and requirements.

**Environment:** SAP BW, SAP BusinessObjects, Lumira, Informatica IDQ, Informatica MDM, Jira

**Education:**

**Master of Science in Software Engineering -** San Jose State University - Dec 2022

**Bachelor of Engineering in Computer Science and Engineering -** Anna University - Apr 2015