# Vigneswari Pappu

# Senior Data Engineer

# Email: XXXXXXX@gmail.com

# Phone: 44xxxxxxxxxxx

# SUMMARY

* 10 + years of experience in Big Data technologies, specializing in Hadoop, Spark, and SQL across both cloud and on-premises environments.
* Expertise in building and optimizing scalable ETL pipelines using Python, Scala, dbt, and Sqoop, integrating data from diverse sources into Hadoop and Snowflake.
* Designed and managed data models using Hive, HBase, and Snowflake to handle large-scale structured and unstructured data, ensuring high performance and efficient storage.
* Extensive experience with real-time data streaming and ingestion using Kafka and Spark, seamlessly integrating with cloud-based platforms like AWS and Google Cloud.
* Automated and orchestrated data workflows with Airflow and Jenkins, ensuring efficient monitoring, scheduling, and management of complex data pipelines.
* Proven track record of ensuring data quality and pipeline performance through continuous testing, validation, and performance monitoring using tools like Pandas, Pytest, and Elasticsearch.
* Implemented robust security controls using IAM, Kerberos, SSL encryption, and data governance policies to ensure data protection and compliance with industry standards.
* Hands-on experience with cloud technologies such as AWS Glue, EMR, and Google Cloud Dataproc, along with managing on-premises Hadoop clusters, delivering scalable and cost-effective data solutions.
* Developed insightful data visualizations using Tableau and Grafana, providing business stakeholders with actionable insights and real-time metrics.
* Skilled in containerization and orchestration with Docker and Kubernetes, ensuring seamless deployment, scalability, and management of big data applications.
* Strong experience in incident management and optimization of data environments through integration with ServiceNow and continuous system monitoring.
* Google Cloud Bigtable for database management and integration with Hadoop.

# SKILLS

* Programming Languages: Spark, Python, Scala, SQL, Oracle SQL, Bash Scripting
* Big Data Frameworks: HDFS, HBase, Hive, Cloudera, Snowflake, Elasticsearch, Parquet, Kafka, Flume, Spark Structured Streaming
* Data Warehousing: S3, Google Cloud Storage, Azure Storage, Redshift, Redshift Spectrum, Athena, Google BigQuery, Snowflake, DBT
* Cloud Technologies: Glue, EMR, AWS Lambda, Step Functions, Cloud watch, Kinesis, DynamoDB, Azure HDInsight, Azure Synapse, Azure Data Factory, Google Cloud Dataproc, IBM Cloud
* Database Management Systems: MySQL, Microsoft SQL Server, Postgre SQL, HBase
* Data Visualization: Tableau, Grafana
* Testing and Automation: Pytest, Scala test, Airflow
* Containerization: Docker, Kubernetes
* Integration and APIs: RESTful APIs
* Continuous Integration and Deployment: Jenkins, AWS Code Pipeline
* Other: SSIS, EC2, IBM Spectrum Scale

# EXPERIENCE

***Senior Big Data Engineer***

***Wren Kitchens, London, UK, February 2023 – Present***

Wren Kitchens is a UK-based kitchen retailer and manufacturer known for its high-quality, bespoke kitchen products and services. The company offers a wide range of kitchen designs, styles, and accessories, catering to various preferences and budgets.

* Successfully sourced missing data from Jira Cloud using Jira API and automated data extraction with AWS Glue.
* Conducted capacity planning and transformed raw data from Jira Cloud into structured formats using AWS Glue for seamless analysis.
* Developed custom data visualizations using OTBI and OAC, providing insights into business data.
* Implemented data cleansing and validation techniques within AWS Glue, ensuring data quality and reliability.
* Managed query execution permissions and access control in Athena and monitored performance with AWS CloudWatch.
* Designed and maintained ETL pipelines to extract, transform, and load data into QlikView and other BI platforms.
* Implemented automated incident management workflows with ServiceNow and proactive anomaly detection using AWS CloudWatch.
* Led data migration efforts using Azure Data Factory and AWS DMS, ensuring seamless migration from on-premises databases to the cloud.
* Integrated Jira API and AWS Glue for seamless data flow and implemented AWS Redshift for data storage and transformation.
* Developed and managed data pipelines within Airbyte and ensured secure data transfer using Amazon S3.
* Established data governance frameworks and collaborated with stakeholders to define data transformation objectives.
* Utilized Debezium for capturing real-time data changes and Postman for API testing, ensuring data accuracy.
* Employed AWS Lambda for serverless execution of key tasks and automation of data workflows.
* Optimized PySpark-based ETL processes for extracting and loading data into a Hadoop data lake.
* Designed data validation routines and utilized AWS Glue Data Catalog for seamless metadata management.
* Collaborated with teams to enhance DBT models for efficient data pipelines and integrated Jenkins with GitLab for automated deployments.
* Created and optimized database schemas in MySQL and Oracle for improved performance and specific use cases.
* Automated data extraction, transformation tasks, and data enrichment using AWS Glue DataBrew and AWS Lambda.

***Senior Big Data Engineer***

***IBM UK, London, UK, April 2022 – February 2023***

IBM UK provides a wide range of technology services and solutions to businesses and organizations in the United Kingdom, including cloud computing, artificial intelligence, analytics, block chain, and security.

* Led the development of a big data platform for IBM UK using Azure services such as Azure Data Factory (ADF), Azure Databricks, Azure Blob Storage, and Event Hubs.
* Designed and implemented ETL pipelines using Azure Databricks and Python for processing and analyzing large datasets stored in Azure Data Lake and Cosmos DB.
* Applied basic data transformations within Azure Data Factory, such as data cleaning, filtering, and aggregation.
* Collaborated with team members using pair programming, leading to a 20% increase in productivity.
* Worked on data preparation tasks within Power BI, cleansing and structuring data to make it suitable for visualization.
* Created interactive data visualizations, graphs, and charts in Power BI to support data-driven decision-making.
* Worked with Azure Storage, Azure SQL, Azure HDInsight, and Azure Synapse for cloud-based big data solutions.
* Developed ETL processes to move data between MySQL and Azure SQL, ensuring data consistency and accuracy.
* Monitored and managed data pipelines using Azure Monitor and Grafana for tracking performance metrics.
* Configured Azure Logic Apps to automate incident management workflows, routing incidents based on predefined criteria.
* Implemented CI/CD processes for complex big data pipelines using Azure DevOps, ensuring smooth integration and automated testing.
* Used Azure Event Hubs for real-time data streaming and analysis.
* Used Azure Data Factory and Airflow for scheduling and monitoring data workflows, automating data pipelines.
* Evaluated and selected modern data storage technologies, such as Azure Data Lake and Cosmos DB, for diverse data types and workloads.
* Implemented security controls with Azure Key Vault and SSL encryption to protect sensitive data and ensure compliance with data privacy regulations.
* Developed and planned agendas for "Show and Tell" sessions, selecting presenters and setting objectives.
* Participated in pair programming with other developers to enhance productivity and code quality.
* Developed geospatial database architectures using PostgreSQL, PostGIS, and Cosmos DB for managing large spatial datasets.
* Built and maintained data warehouse instances for efficient data storage and analysis.
* Managed relational databases using Azure SQL and MySQL for integration with Azure services.
* Developed PySpark solutions on Azure Databricks for big data processing.
* Managed data transformations and schema migrations using dbt in an Azure SQL Data Warehouse environment.
* Used Azure Data Lake and Data Catalog for data lineage and metadata management.
* Designed complex SQL queries for ETL processes, improving data quality and performance.
* Utilized Debezium for change data capture across MySQL, PostgreSQL, and Cosmos DB databases.
* Developed custom API connectors and data pipelines to integrate with RESTful APIs and other external systems.
* Integrated Azure Monitor with Azure Functions to trigger alerts and automated tasks based on system performance metrics.
* Created dashboards and reports in Power BI and Tableau to provide business insights to stakeholders.
* Embraced Agile practices and methodologies to accommodate changes and feedback within the sprint framework.
* Managed Azure resources, ensuring cost optimization and adherence to best practices for cloud governance.
* Developed Spark-SQL and Streaming code in Scala and PySpark for big data processing in Azure Databricks.
* Monitored Agile metrics like velocity and cycle time to drive team improvements in productivity.
* Collaborated with the DevOps team to ensure high availability and scalability of the Azure-based big data platform.

## 

## *Big Data Engineer*

***ASOS, London, UK -, February 2021 - April 2022***

ASOS is a British online fashion and beauty retailer that specializes in fast fashion and operates in over 240 countries. The company was founded in 2000 and has since become one of the world's leading online fashion retailers, offering a wide range of clothing, shoes, accessories, and beauty products for men and women.

* Wrote complex SQL queries for an on premises PostgreSQL database as well as an AWS Redshift Data Warehouse
* Collaborated with data scientists to leverage AWS Athena for interactive data exploration and analysis.
* Implemented cost-control measures by optimizing AWS Athena query execution and monitoring usage.
* Worked with AWS S3 for cloud-based big data solutions.
* Used AWS Glue for ETL and data transformation in the cloud.
* Created and optimized Hive tables and queries to support ad hoc data analysis and reporting
* Transformed data as it moves between MySQL and Oracle databases, ensuring data compatibility and quality.
* Imported unstructured data into the Hadoop Distributed File System (HDFS using Spark Streaming and Kafka, and wrote complex Hive queries, Spark SQL queries and UDFs
* Proficiently used Agile project management and collaboration tools such as Jira, Trello, or Asana to plan, track, and report on project progress.
* Using DBT to write and manage SQL queries that transform raw data into structured and usable formats.
* Leveraged Jenkins to automate the deployment and orchestration of Apache Spark and Hadoop clusters, optimizing resource utilization and reducing deployment time by 30%.
* Spearheaded the migration of legacy data formats to modern, standardized data structures, improving data accessibility and compatibility.
* Identified and resolved performance issues in both MySQL and Oracle databases by fine-tuning configurations and indexes.
* Built a system for geospatial data ingestion, cleaning, and transformation that uses Apache NiFi and other ETL tools to automate the data preparation process.
* Created RDDs and DFs for the required input data and performed the data transformations using Spark Python.
* Conducted pair programming sessions to teach junior team members new programming skills and technologies, improving their proficiency and confidence.
* Developed Spark SQL queries and Data Frames and imported data from Data sources,
* Integrate Kafka and Spark using Avro for serializing and de-serializing data, and for Kafka Producer and Consumer.
* Used Docker and Kubernetes for containerization and orchestration of data platform.
* Developed custom data connectors to integrate with RESTful APIs and other external systems.
* Used Azure Data Factory for data integration and management in the cloud.
* Devised ETL functions between Oracle and Amazon Redshift. Used Rest API to Access HBase data to perform analytics.
* Conducted data profiling and analysis to identify data quality issues and improve data accuracy.
* Developed dashboards and reports using Tableau for visualizing and presenting data insights to stakeholders.
* Used Kafka for real-time data ingestion and processing and integrated it with Hadoop platform.
* Implemented Parquet file format for optimized data storage and retrieval in Hadoop.
* Designed and implemented a system for geospatial data caching and retrieval using technologies like Redis and Apache Cassandra to optimize query performance.
* Implemented Spark Structured Streaming for real-time data processing and analysis..
* Built data pipelines using Cassandra and Spark to process and analyze large volumes of data in real-time, enabling data-driven insights and decision-making.
* Implemented data modeling best practices in Cassandra to optimize query performance and reduce data duplication.
* Calculated and interpreted sprint velocity, which provided insights into the team's capacity and informed future sprint planning.
* Used Elastic search, Log stash, and Kibana for log analysis and monitoring of system performance.
* Used Snowflake for cloud-based data warehousing and data analytics.

***Big Data Engineer***

***Travis Perkins | Northampton |, UK, April 2020 - February 2021***

EDF Energy is a British energy company that provides gas and electricity to homes and businesses in the United Kingdom. The company is a subsidiary of EDF Group, a French multinational energy company, and was founded in 2002 following the acquisition of London Electricity.

Travis Perkins is a leading supplier of building materials to the trade, construction, and home improvement markets in the United Kingdom. The company serves thousands of customers across its nationwide branches, providing essential products for both small-scale and large-scale construction projects.

* Used Snowflake and Google Big Query for cloud-based data warehousing and data analytics.
* Used Pandas and Pytest for automated testing of data pipelines and data quality.
* Designed and delivered technical training sessions for junior team members, improving their knowledge and skills in big data technologies and methodologies.
* Coached junior team members in developing scalable and efficient big data solutions using Hadoop and Apache Spark technologies.
* Implementing automation processes using DBT to streamline data transformation and modeling tasks.
* Created Hive Generic UDF's to process business logic that varies based on policy.
* Used Elastic search for log analysis and monitoring of system performance.
* Used Jenkins for continuous integration and deployment of data pipelines.
* Created Docker containers to deploy and manage data pipelines
* Managed the sprint backlog, including organizing, prioritizing, and assigning user stories or tasks to team members.
* Developed data transformation and enrichment workflows using Azure Data bricks, improving overall system performance and efficiency.
* Documenting DBT models and transformations for clarity and knowledge sharing.
* Conducted performance tuning and optimization of Cassandra clusters, improving query response times and reducing resource utilization.
* Assisted in the training and onboarding of team members and stakeholders in Agile principles and practices.
* Designed and implemented a hybrid cloud architecture for Cassandra clusters, enabling seamless data replication and access across multiple cloud providers and on-premises systems.
* Implemented data quality and validation checks using Azure Data Factory and Azure Synapse
* Developed advanced incident reports in Service Now, providing detailed metrics on incident resolution times, enabling data-driven improvements in our incident management processes.
* Used Kubernetes for container orchestration and scaling of data platform.
* Used Kafka for real-time data ingestion and processing and integrated it with Spark platform..
* Developed a geospatial data processing platform using Spark Streaming, Spark Structured Streaming, and PySpark to ingest and analyze real-time geospatial data from various sources.
* Built a scalable data storage solution using Azure Storage to store and access large datasets in the cloud.
* Used Google Cloud Dataproc for processing and analyzing large datasets in the cloud.
* Used Google Big Query for ad-hoc and interactive analysis of data.
* Developed custom data connectors to integrate with RESTful APIs and other external systems.
* Conducted data profiling and analysis to identify data quality issues and improve data accuracy.
* Worked with Cloudera for managing big data solutions and integration with Google Cloud services.
* Implemented security controls using Google Cloud security services and SSL to protect sensitive data and comply with data privacy regulations.
* Developed ETL jobs using Azure Data Factory and Spark to automate and orchestrate data processing and analysis workflows.
* Developed dashboards and reports using Tableau for visualizing and presenting data insights to stakeholders.
* Built data processing workflows using Azure Functions to enable server less computing for data processing and analysis.
* Designed and developed data pipelines in an Azure environment using ADL Gen2, Blob Storage, ADF, Azure Data bricks, Azure SQL and Azure Synapse for analytics and MS Power BI for reporting.
* Implemented Parquet file format for optimized data storage and retrieval in HDFS.
* Conducted performance and load testing of the big data platform using JMeter and other tools.
* Used Google Cloud Storage for data backup and archiving.
* Provided guidance and support to junior developers in troubleshooting and resolving complex issues related to big data processing and storage.
* Worked with Azure Synapse for cloud-based data warehousing and data analytics.
* Used Google Cloud Pub/Sub for real-time messaging and data streaming.
* Used Google Cloud Big table for NoSQL database management and integration with Hadoop.

***Big Data Engineer***

***HSBC, London, UK, May 2019 - March 2020***

HSBC is a British multinational banking and financial services company, with headquarters in London. The company was founded in 1865 to facilitate trade between Europe and Asia, and has since grown to become one of the largest banking and financial services organizations in the world.

* Developed and implemented a real-time data pipeline using Kafka, Spark, and HDFS to enable near-instantaneous data processing and analysis
* Designed and built data pipelines using Python, Scala, and PySpark for processing large volumes of financial data.
* Utilized HBase to store and retrieve data, and implemented data modeling techniques to optimize data access and retrieval.
* Utilized Oozie workflows to run Hive and Impala jobs in Hadoop ecosystem for data processing and analytics.
* Created a data warehouse using Hive and Hadoop for efficient storage and retrieval of large volumes of financial data.
* Developed ETL pipelines using Sqoop and Airflow to extract data from various sources, transform it, and load it into HDFS.
* Integrated Cloudera Manager and Zookeeper to automate and streamline the management of the Hadoop cluster.
* Implemented Flume to collect and aggregate log data from various sources and store it in HDFS.
* Developed and maintained automated tests using Pytest to ensure data quality and consistency.
* Built and maintained a RESTful API for accessing data from the Hadoop cluster using Python and Flask.
* Developed and maintained dashboards using Tableau for visualization of financial data.
* Utilized PostgreSQL to store metadata and track data lineage.
* Monitored and managed the Hadoop cluster using Splunk for log analysis and monitoring.
* Created data pipelines using Oozie to automate and schedule data processing and analysis.
* Utilized Jenkins to automate build and deployment processes.
* Developed and implemented data retention policies to manage the lifecycle of data stored in Hadoop.
* Implemented Kerberos authentication to secure the Hadoop cluster.
* Conducted performance tuning and optimization of the Hadoop cluster to ensure optimal performance and efficiency.
* Developed and implemented disaster recovery plans and procedures to ensure business continuity in the event of a data center outage.
* Conducted data analysis to identify trends and patterns in financial data

***Big Data Engineer***

***Rolls Royce Holdings, London, UK, April 2018 - May 2019***

Rolls Royce Holdings is a world-renowned engineering company that designs, manufactures, and distributes power systems for aviation, marine, and industrial applications. With a strong focus on innovation and technology, Rolls Royce plays a critical role in the aerospace industry, supplying engines and power systems globally.

* Developed and implemented a data ingestion pipeline using Kafka, Spark, and Hadoop to process high volume of data from various sources.
* Designed and built a data warehouse using Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse) for efficient storage and retrieval of large volumes of data.
* Implemented ETL pipelines using Azure Data Factory and Python to extract data from various sources, transform it, and load it into Azure Synapse.
* Implemented data partitioning and sharding to optimize data retrieval and processing.
* Utilized Spark for processing and analyzing large volumes of data.
* Developed and maintained automated tests using Pytest to ensure data quality and consistency.
* Conducted performance tuning and optimization of the data warehouse to ensure optimal performance and efficiency.
* Developed and maintained dashboards and reports using Tableau for visualization of data.
* Built and maintained a RESTful API for accessing data from the data warehouse using Python and Flask.
* Implemented Bash scripting for automation of various processes.
* Utilized Jenkins to automate build and deployment processes.
* Implemented Docker and Kubernetes for containerization and orchestration of data processing jobs.
* Utilized Azure Storage for storing and retrieving data.
* Utilized Azure SQL Database for managing relational data.
* Implemented Microsoft SQL Server for efficient storage and retrieval of large volumes of data.
* Utilized SSIS for developing ETL packages and workflows.
* Utilized Airflow for scheduling and monitoring data processing workflows.

***Big Data Developer***

***Bloomberg, London, UK, January 2017 - March 2018***

Bloomberg is a global financial services and media company that operates in the United Kingdom, with a significant presence in London. Bloomberg is known for its Bloomberg Terminal, a data and analytics platform widely used by financial professionals and institutions worldwide.

* Installed and configured Hadoop clusters on both development and production environments.
* Created Flume configurations to ingest real-time log data from web servers into HDFS.
* Developed custom Hive scripts for data transformation and reporting purposes.
* Implemented Pig Latin scripts to clean and preprocess raw data before storage in HDFS.
* Designed and optimized HBase data models for efficient read and write operations.
* Managed cluster resources effectively using YARN to ensure fair resource allocation.
* Developed and debugged Map Reduce applications in Java to process large datasets.
* Set up monitoring tools like Nagios and Ganglia to track cluster performance and resource usage.
* Utilized data compression techniques (Snappy, LZO) to reduce storage space.
* Implemented data backup and recovery strategies, including HDFS snapshots.
* Enhanced cluster security by configuring Kerberos authentication and authorization.
* Conducted data quality checks and wrote custom validation scripts to maintain data integrity.
* Created Oozie workflows for scheduling and orchestrating data processing jobs.
* Performed capacity planning by analyzing data growth trends and optimizing HDFS storage.
* Integrated various components of the Hadoop ecosystem, including Hive, Pig, and HBase, to create end-to-end data pipelines.

***Software Developer***

***LILT Solution, Coventry, UK, May 2014 - December 2016***

LILT Solution is a leading software development company specializing in building enterprise-level applications and data solutions for various industries. During my time at LILT, I worked on various data processing and analysis projects, leveraging a mix of Python, Oracle, MySQL, Spark, and Hadoop.

* Developed and deployed a Flask-based backend to interface with OpenAI's API, generating dynamic metadata used to configure and display widgets in a React frontend.
* Designed and implemented React components that dynamically render widgets based on metadata received from the Flask backend, enhancing user interaction and experience.
* Implemented caching strategies using Redis (or another caching solution) to store frequently accessed metadata and API responses, significantly improving application performance and reducing server load
* Used **PySpark** for processing and analyzing large data volumes, delivering results to downstream applications.
* Integrated **Apache Kafka** to handle real-time data streaming, using **Python** to write custom consumers that processed streaming data and stored outputs in **HDFS** or **HBase** for long-term storage and querying.
* Developed and optimized data processing pipelines using **Hive**, **Impala**, and **HBase**. Designed data models for **HBase** to handle large-scale reads and writes efficiently.
* Implemented caching strategies using **Redis** to store frequently accessed metadata and API responses, which significantly improved the performance of the Flask-based application and reduced server load.
* Applied **Snappy** and **LZO** compression techniques in Hadoop to reduce storage costs while maintaining fast data access.
* Configured **Kerberos** authentication for the Hadoop cluster, ensuring secure access to critical data pipelines and services like **Hive**, **Kafka**, and **HBase**.
* Developed Python-based **ETL** pipelines to ingest data from **Oracle** and **MySQL** databases into **HDFS** for further processing, automating tasks using **Oozie** workflows.
* Utilized tools such as **Nagios** and **Ganglia** to monitor cluster performance, with additional custom monitoring dashboards built using **Flask** to display real-time system health metrics.

# EDUCATION

Bachelor of Technology (B.Tech) in Computer Science and Engineering Jawaharlal Nehru Technological University 2010 -2014