**Name: Supriya Parvatham**

**Email ID:**  [**supriyap5092@gmail.com**](mailto:supriyap5092@gmail.com)

**Phone:** (747) 999-6694

**Senior Big Data Engineer**

**PROFILE SUMMARY**

* Over 9+ years of experience designing and implementing various big data solutions.
* Strong experience in both Waterfall and Agile methodologies.
* Expertise in using major components of the Hadoop ecosystem, including HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Flume, Oozie, and Zookeeper.
* Knowledge of ETL methods for data extraction, transformation, and loading in corporate-wide ETL solutions and data warehouse tools.
* Skills in data modeling, data mining, and improving data reliability, efficiency, and quality.
* Hands-on experience working with various file formats like delimited text files, clickstream log files, Apache log files, Avro files, JSON files, and XML files, and columnar file formats like RC, ORC, and Parquet formats.
* Knowledge of NoSQL databases like HBase, Azure, MongoDB, and Cassandra.
* Proficiency in Databricks MLflow for running machine learning models on distributed platforms.
* Strong knowledge of architecture and components of Spark, including Spark Core, SparkSQL, and Spark Streaming.
* Familiarity with Spark Scala's functional style programming techniques.
* Experience working with Kafka and Kafka brokers for processing live streaming data.
* Experience in reading multiple data formats on HDFS using Scala, creating DataFrames by loading data from Hive tables, and storing prep data in AWS S3.
* Experienced in designing and implementing enterprise data warehouse, business intelligence, analytical, batch/real-time/near-real-time streaming big data solutions.
* Proficient in both Waterfall and Agile methodologies.
* Expertise in major components of the Hadoop ecosystem, including HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Flume, Oozie, and Zookeeper.
* Strong understanding of distributed systems, HDFS architecture, and internal working details of MapReduce and Spark processing frameworks.
* Knowledge of ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
* Skilled in data modelling, data mining, and improving data reliability, efficiency and quality.
* Experienced in importing and exporting data using Sqoop from HDFS to Relational Database systems and vice versa and loading into Hive tables, which are partitioned.
* Deep understanding of Hive, Pig, and Sqoop to write MapReduce jobs for data analysis, write custom UDFs, and optimize Hive Queries.
* Proficient in working with various file formats like delimited text files, click stream log files, Apache log files, Avro files, JSON files, and XML Files, and using different columnar file formats like RC, ORC, and Parquet formats.
* Knowledge of NoSQL databases like HBase, Azure, MongoDB, and Cassandra.
* Experience in running queries using Impala and using BI tools to run ad-hoc queries directly on Hadoop.
* Proficient in Databricks MLflow for running machining learning models on distributed platforms.
* Experience in extracting files from MongoDB through Sqoop and placing them in HDFS and processing.
* Worked with NoSQL databases like HBase in creating HBase tables to load large sets of semi-structured data coming from various sources.
* Strong knowledge of architecture and components of Spark, including Spark Core, SparkSQL, and Spark Streaming.
* Familiar with Spark Scala's functional style programming techniques like Anonymous Functions (Closures), Higher Order Functions and Pattern Matching.
* Experience in converting Hive/SQL queries into Spark transformations using Spark Data frames and Scala.
* Experience in working with Kafka and Kafka brokers to initiate spark context and processing live streaming, developing custom Kafka producer and consumer for different publishing and subscribing to Kafka topics.
* Good working experience on Spark (spark streaming, spark SQL) with Scala and Kafka.
* Worked on reading multiple data formats on HDFS using Scala, creating Data frames by loading data from Hive tables, and creating prep data and stored it in AWS S3

**TECHNICAL PROFICIENCY**

| **Big Data Technologies** | Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Flink, Presto, Druid, Delta Lake, Apache Beam, Apache Nifi, Apache NiFi Registry, Apache Airflow, Apache Druid, Apache Pulsar, AWS Kinesis, Google Cloud Dataflow, Google BigQuery, Databricks Delta, Snowflake, Azure Data Lake, Azure Stream Analytics |
| --- | --- |
| **Hadoop Distribution** | Cloudera CDP, Hortonworks HDF, Apache Hadoop, Amazon EMR, Microsoft HDInsight, Google Cloud Dataproc |
| **Programming Languages** | SQL, PL/SQL, Python, R, PYSpark, Pig, Hive QL, Scala, Shell Scripting, Regular Expressions, Java, Kotlin, Go, Rust |
| **Spark components** | RDD, Spark SQL (Data Frames and Dataset), Spark Streaming, Spark MLlib, Spark GraphX, SparkR, PySpark |
| **Cloud Infrastructure** | AWS, Azure, GCP, Alibaba Cloud, IBM Cloud, Oracle Cloud |
| **Databases** | Oracle, Teradata, MySQL, SQL Server, NoSQL Databases (HBase, MongoDB, Cassandra, Couchbase, DynamoDB) |
| **Scripting & Query Languages** | Shell scripting, SQL, GraphQL, Apache Drill |
| **Version Control** | CVS, SVN, Clear Case, GIT, Bitbucket, GitHub, GitLab |
| **Build Tools** | Maven, SBT, Gradle |
| **Containerization Tools** | Kubernetes, Docker, Docker Swarm, Apache Mesos, Apache Aurora |
| **Reporting Tools** | JUnit, Eclipse, Visual Studio, NetBeans, Azure Databricks, UNIX, Power BI, SAS, Tableau, Looker, Mode Analytics, QuickSight |

**WORK EXPERIENCE:**

**Kollabio, Ashburn, VA Feb 2022- Till Now**

**Title: Sr. Data Engineer**

Roles & Responsibilities:

* Design, develop and maintain complex data pipelines using the latest tools and technologies such as AWS Glue, Apache Airflow, and Google Cloud Dataflow for data ingestion, transformation, and storage
* Utilize AWS, and Google Cloud services to extract, transform, and load data from various sources into a data warehouse or data lake for downstream processing.
* Develop and maintain data models, data dictionaries, and data lineage using tools such as ERWIN, DBT, and Alation.
* Skilled in optimizing Druid performance by tuning segment size and number, caching, indexing, and other techniques.
* Experience in testing GraphQL APIs and debugging issues. Highlight any testing frameworks or tools you have used to ensure the quality and reliability of your GraphQL implementations
* Use SQL, Spark SQL, and NoSQL databases like MongoDB and Cassandra for data analysis and querying.
* Develop and maintain data pipelines using programming languages like Python, Scala, and Java, and frameworks such as PySpark, Spark, and Hadoop.
* Proficient in implementing data privacy measures and ensuring compliance with relevant regulations such as GDPR, CCPA, and HIPAA.
* Understanding and hands-on experience with GraphQL, including its core concepts, schema definition, and query language.
* Software development involving cloud computing platforms like Amazon Web Services (AWS), and Google Cloud (GCP).
* Experienced in working with Amazon Web Services (AWS) using S3, EMR, Redshift, Athena, Glue Meta store etc.,
* Experience in GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Data Proc, Storage, Composer and Stack driver.
* Developed and implemented workflows using Azure Logic Apps to automate business processes and integrate systems across different platforms.
* Created visually appealing and intuitive workflows using the Azure Logic Apps Designer, leveraging its drag-and-drop interface and pre-built connectors.
* Extensive experience developing serverless functions in Azure using Azure Functions, creating event-driven microservices, and integrating them into Azure Logic Apps workflows.
* Experience in designing and implementing data anonymization and pseudonymization techniques to protect sensitive information while maintaining data utility.
* Skilled in integrating privacy-enhancing technologies (PETs) such as differential privacy, secure multi-party computation, and homomorphic encryption into big data pipelines.
* Implement data encryption and data security using Hashing, encryption algorithms, and AWS KMS.
* Work with machine learning engineers and data scientists to build and maintain machine learning models using tools like TensorFlow, Keras, and Scikit-Learn.
* Build and maintain data visualization dashboards using tools such as Tableau, PowerBI, and Looker for real-time data analysis and insights.
* Developed and implemented Azure Cosmos DB solutions for multiple clients, ensuring high availability, scalability, and optimal performance.
* Designed and implemented data models using Azure Cosmos DB, leveraging different APIs such as SQL, Gremlin, MongoDB, and Cassandra.
* Collaborated with cross-functional teams to gather requirements, design database architectures, and optimize query performance.
* Develop and implement data governance policies and procedures to ensure data quality, accuracy, and compliance with regulatory requirements.
* Work with DevOps engineers to deploy and manage data pipelines and applications in a cloud-based infrastructure using tools such as Kubernetes, Docker, and Jenkins.
* Collaborate with cross-functional teams to gather requirements, design, and develop end-to-end data solutions.
* Solid expertise in working with PostgreSQL, leveraging its advanced features for efficient data storage, retrieval, and management.
* Perform data profiling, data cleansing, and data validation using tools like Trifacta, Talend, and Informatica.
* Develop and maintain ETL processes using tools such as SSIS, Informatica, and Talend.
* Perform performance tuning and optimization of SQL queries and data pipelines using tools such as Apache Spark, Presto, Hive, and Impala.
* Develop and maintain real-time data streaming pipelines using tools such as Kafka, Kinesis, and Spark Streaming.

**Environment:** AWS, Google Cloud, SQL Server, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), Databricks, SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Spark, Kafka, Impala, Python, Power BI, Tableau, Presto, Hive/Hadoop, Snowflakes, TensorFlow, Keras, Scikit-Learn, Kubernetes, Docker, Jenkins, Trifacta, Talend, Informatica, Snowflakes.

**USAA, San Antonio, TX June 2019 – Jan 2022**

**Title: Sr. Big Data Engineer**

**Roles & Responsibilities:**

* Developing and implementing end-to-end data engineering solutions using tools such as Apache Kafka, Apache NiFi, and Apache Airflow to create data pipelines for real-time data streaming and batch processing.
* Designing and implementing data models and schemas using NoSQL databases such as MongoDB, Cassandra,and DynamoDB.
* Building and maintaining cloud-based data lakes using AWS S3 and Azure Data Lake Storage Gen2, and processing the data using tools such as AWS Glue, Azure Databricks, and HDInsight.
* Implementing data governance and data quality processes using tools such as Apache Atlas, Cloudera Navigator, and Collibra.
* GraphQL with various data sources, such as databases, REST APIs, or microservices.
* Creating and managing data pipelines using tools such as Apache Beam, Apache Flink, and Apache Spark Streaming for real-time data processing.
* Strong understanding of data governance frameworks and practices, including data classification, access controls, and data retention policies.
* Familiarity with privacy impact assessments (PIAs) and conducting privacy audits to identify and mitigate potential privacy risks in big data systems.
* Hands-on experience in deploying and managing data privacy tools and platforms, such as data masking, tokenization, and data loss prevention (DLP) solutions.
* Developing and implementing machine learning models using tools such as TensorFlow, Keras,and
* PyTorch for predictive analytics and natural language processing (NLP).
* Implementing big data security solutions using tools such as Apache Ranger, Cloudera Sentry, and AWS Identity and Access Management (IAM).
* Developing and implementing data visualization solutions using tools such as Tableau, Power BI, and Apache Superset for reporting and dashboarding.
* Utilizing containerization technologies such as Docker and Kubernetes to deploy and manage big data applications in a scalable and efficient manner.
* Developing and implementing data governance policies and best practices to ensure compliance with data privacy regulations such as GDPR and CCPA.
* GraphQL query patterns and optimize them for better performance and resource utilization
* Developing and implementing disaster recovery and business continuity plans for big data systems using tools such as Apache Hadoop Distributed File System (HDFS) and AWS Elastic Block Store (EBS).
* Collaborating with cross-functional teams such as data scientists, business analysts, and software developers to deliver high-quality big data solutions.
* Implementing DevOps practices for big data engineering projects using tools such as Jenkins, GitLab, and GitHub for continuous integration and delivery (CI/CD).
* Designing and implementing data cataloging solutions using tools such as Apache Atlas, AWS Glue Data Catalog, and Collibra Catalog.

**Environment**: Hadoop, Kafka, Spark, Sqoop, Spark SQL, Spark-Streaming, Hive, Impala, Scala, pig, NoSQL, Oozie, Hbase, Data Lake, Python , Azure, Databricks, AWS(Glue, Lambda, StepFunctions, SQS, Code Build, Code Pipeline, EventBridge, Athena), Unix/Linux Shell Scripting, Informatica PowerCenter.

**Paradigm, San Diego, CA Dec 2017- May 2019**

**Title: Big Data Engineer**

**Roles & Responsibilities:**

* Experience in Job management using Fair scheduler and Developed job processing scripts using Apache Airflow workflow.
* Used Spark, Hive, and Databricks for implementing the transformations needed to join the daily ingested data to historic data.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time.
* Developed Spark scripts by using Scala shell commands and PySpark as per the requirement.
* Used Spark API over EMR Cluster Hadoop YARN to perform analytics on data in Hive.
* Developed Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD in Spark for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Writing PySpark and spark SQL transformation in Azure Databricks to perform complex transformations for business rule implementation.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, and Pair RDD's.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark.
* Developed logistic regression models (using R programming and Python) to predict subscription response rate based on customer’s variables like past transactions, response to prior mailings, promotions, demographics, interests and hobbies, etc.
* Created Tableau dashboards/reports for data visualization, Reporting and Analysis and presented it to Business.
* Created/ Managed Groups, Workbooks and Projects, Database Views, Data Sources, and Data Connections.
* Worked with the Business development managers and other team members on report requirements based on existing reports/dashboards, timelines, testing, and technical delivery.
* Knowledge in Tableau Administration Tool for Configuration, adding users, managing licenses and data connections, scheduling tasks, embedding views by integrating with other platforms.
* Developed dimensions and fact tables for data marts like Monthly Summary, Inventory data marts with various Dimensions like Time, Services, Customers, and policies.
* Developed reusable transformations to load data from flat files and other data sources to the Data Warehouse.
* Assisted operation support team for transactional data loads in developing SQL Loader & Unix scripts.
* Implemented Python script to call the Cassandra Rest API, performed transformations and loaded the data into Hive.
* Extensively worked on Python and build the custom ingest framework.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations, and other during the ingestion process itself.
* Experienced in writing live Real-time Processing using Spark Streaming with Kafka.
* Created Cassandra tables to store various data formats of data coming from different sources.
* Designed, developed data integration programs in a Hadoop environment with NoSQL data store Cassandra for data access and analysis.
* Generated Custom SQL to verify the dependency for the daily, Weekly, Monthly jobs.
* Using Nebula Metadata, registered Business and Technical Datasets for corresponding SQL scripts
* Experienced in working with spark ecosystem using Spark SQL and Scala queries on different formats like text file, CSV file.
* Developed spark code and spark-SQL/streaming for faster testing and processing of data.
* Closely involved in scheduling Daily, Monthly jobs with Precondition/Postcondition based on the requirement.
* Monitor the Daily, Weekly, Monthly jobs and provide support in case of failures/issues.

Environment: Hadoop YARN, Azure, Databricks, Spark 3.0, Spark Streaming, Spark SQL, Scala, Kafka, Python, Impala.

**KPIT Technologies, India Feb 2016 - Oct 2017**

**Title: Data Engineer**

**Roles & Responsibilities:**

* Built and architected end-to-end ETL pipelines using GCP data services such as Cloud Dataflow, Cloud Storage, and BigQuery.
* Proficient in AWS components such as EC2, S3, and implemented Continuous Delivery pipelines using Docker and GitHub.
* Designed and developed serverless data ingestion pipelines using GCP Cloud Functions and Python to load data into BigQuery from on arrival CSV files in GCS bucket.
* Used Apache Kafka for data streaming and integrated with Hadoop ecosystem components for real-time data processing.
* Implemented data quality and data profiling frameworks using Impala, Apache Atlas, and Apache NiFi to ensure data accuracy, completeness, and consistency.
* Worked on GCP and AWS cloud environments to develop and deploy scalable data pipelines and machine learning back-end pipelines using Apache Airflow, Kubernetes, and Docker.
* Utilized Pandas, NumPy, and Matplotlib libraries for data processing, analysis, and visualization.
* Expertise in Hadoop ecosystem components such as Hadoop Distributed File System (HDFS), Apache Hive, and Apache Spark for storing, processing, and analyzing big data.
* Experience in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling, and Slowly Changing Dimensions (SCD) in data warehousing.
* Proficient in statistical modeling and machine learning techniques such as Decision Trees, Linear/Logistic Regressors, and implemented machine learning models on Spark and Hadoop clusters.
* Experience in developing PL/SQL Stored Procedures, Functions, Triggers, Views, and Packages for database management systems and optimization of query performance.
* Worked with Apache NiFi, Apache Flink, and Apache Beam for processing real-time data streams and generating insights and alerts.
* Proficient in developing microservices architectures using containerization and orchestration technologies such as Docker, Kubernetes, and Apache Mesos to manage and scale big data clusters.

**Environment**: Gcp, Bigquery, Gcs Bucket, , Cloud Shell, Hadoop, Spark, Docker, Kubernetes, AWS, Apache Airflow, Python, Pandas, Matplotlib, seaborn library, Numpy, ETL workflows, Python, impala, Kafka, Scala.

**CMC Limited ,India Jan 2014- Jan 2016**

**Title: Spark Developer**

**Roles & Responsibilities:**

* Developing and implementing scalable and efficient data pipelines using Apache Kafka, Apache NiFi, and Apache Flume to ingest streaming and batch data into the Hadoop ecosystem.
* Designing and optimizing data storage and processing solutions using cloud-based data services such as Amazon EMR, Amazon Redshift, and Google BigQuery.
* Developing and deploying machine learning models using distributed computing frameworks such as TensorFlow, PyTorch, and Apache MXNet on top of Hadoop/Spark clusters.
* Implementing data governance and security policies using Apache Ranger, Apache Atlas, and Apache Knox to ensure data privacy, compliance, and regulatory requirements.
* Designing and implementing data lakes using Hadoop Distributed File System (HDFS) and Object Storage technologies such as Amazon S3, Google Cloud Storage, and Azure Blob Storage.
* Using Apache Spark Structured Streaming, Apache Flink, and Apache Beam to process real-time data streams and generate insights and alerts.
* Designing and implementing data pipelines using Apache Airflow, Luigi, and Azkaban to orchestrate workflows and automate data processing tasks.
* Implementing microservices architectures using containerization and orchestration technologies such as Docker, Kubernetes, and Apache Mesos to manage and scale Hadoop/Spark clusters.
* Designing and implementing data visualization and reporting solutions using Apache Zeppelin, Tableau, and Power BI to provide interactive and actionable insights to stakeholders.
* Implementing data quality and data profiling frameworks using Apache Atlas, Apache Metron, and Apache NiFi to ensure data accuracy, completeness, and consistency.
* Developed and maintained real-time streaming data pipelines using Apache Kafka and Spark Streaming for near-real-time analytics.
* Worked with Apache NiFi for data ingestion, processing, and distribution across different data sources.
* Created and managed Spark clusters using Kubernetes for better resource allocation and job scheduling.
* Implemented Delta Lake for transactional and ACID-compliant data management and processing using Spark SQL.
* Worked with Apache Airflow for workflow automation and orchestration of data processing pipelines.
* Created machine learning models using Spark MLlib and integrated them into the data processing pipeline for predictive analytics.
* Used Apache Zeppelin for interactive data exploration and visualization with SQL, Python, and Spark.
* Designed and implemented data security and access controls using Kerberos and LDAP for Hadoop clusters.
* Deployed Hadoop clusters on cloud platforms such as Amazon EMR, Google Cloud Dataproc, and Microsoft Azure HDInsight.
* Implemented containerization using Docker and Kubernetes for easier deployment and management of Hadoop and Spark clusters.

**Environments**: Hadoop, HDFS, Spark, Hive, Pig, Sqoop, Oozie, DB2, Java, Python, Oracle, Sql, Splunk, UNIX, Shell Scripting, Apache Kafka, Apache NiFi, Apache Flume, Amazon EMR, Amazon Redshift, Google BigQuery, TensorFlow, PyTorch, Apache MXNet, Apache Ranger.