|  |
| --- |
| -VIJAY KUMAR  **Data Engineer** |

|  |  |  |
| --- | --- | --- |
| **+1 469 846 8345** | [saisri@mastroservices.com](mailto:saisri@mastroservices.com) |  |

Experienced and motivated professional in a Big Data Engineer/Cloud Engineer around 10 years of expertise optimizing data processing through Python applications in Hadoop frameworks. Led ETL Processes, accelerating Campaign data migration into AWS Redshift. Proficient in managing AWS, Azure, GCP accounts, enhancing security, and reducing vulnerabilities. Accomplished in configuring Chef and Cloud Formation for a 25% efficiency gain in service deployment. Possess in- depth ability to gain credibility and build relationships with all levels in an organization, from technical experts to executives and stake holders. Strategically think about business, product, and technical challenges. I am a fast learner adaptable to different work environments and skilled at solving technical issues within projects.

Skills

|  |  |
| --- | --- |
| * Programming languages: C++, Java, Python, R, Spark, Rust, Hadoop, OOP's, Kafka, Scala, Alteryx, Kubernetes, Airflow | * Database: MYSQL, MongoDB, Oracle, DynamoDB, Postgres, Snowflake, PostgreSQL, Redis. |
| * Cloud Services: GCP, Azure, AWS cloud, S3, Step Functions, Neo4j, REL AI, Redshift, Glue. | * Web Technologies: HTML, CSS, JavaScript, TypeScript, Node.js. React, Groovy, Django. |
| * Bigdata Services: HDFS, Spark, Scala, Spark, Map Reduce, HDFS, Oozie, Hive, Pig, Sqoop, Flume, Zookeeper, Kafka and HBase, Databricks, data lake. | * Methodologies: Agile, Waterfall. * Neo4j, Graph QL * Splunk, Flunk |
| * ETL Tools: Pentaho, NIFI, Hyper flow's, AWS Glue, Power BI, Talend, Tableau, SSIS/SSRS | * Scripting Technologies: PowerShell, NodeJS, JavaScript, HTML, Jinga, CSS, Shell Scripting. |
| * Networking protocols: DNS, VPC, TCP/IP, HTTP. | * Version Tools: GitHub, SVN, Visual Studio, MS Excel. |
| * Operating system: Windows Services, Unix, Linux. CentOS. | * ETL development |
| * Data Modeling | * NoSQL Databases |
| * API Development | * Continuous Improvement |
| * Scala Programming | * Relational databases |
| * SQL Programming | * Data Visualization |
| * Python Programming | * Kafka Streaming |
| * Risk Analysis | * Performance Tuning |
| * Advanced SQL | * Excellent Communication |
| * Hadoop Ecosystem | * Data Pipeline Design |
| * RDBMS | * Metadata Management |
| * Big Data Processing | * Data Warehousing |
| * SQL and Databases | * Data Analysis |
| * NIFI |  |

Experience

|  |  |
| --- | --- |
| **Senior Data Engineer | Cloud Engineer**  Capital One | **Dallas**  January 2020 - Present |

* Strengthened AWS Accounts with varied VPCs, enhancing security protocol for Prod and Non-Prod Environments
* Designed and built strong and scalable data pipelines using Spark, RDBMS, NoSQL, Spring Boot, Cosmos DB, and Cassandra
* Orchestrated seamless data flow to Snowflake, Kafka, Hadoop, and SQL for easy access and analysis
* Optimized Python applications with Hadoop, PySpark, and AWS Cloud to build efficient data transformations
* Designed ETL Processes in AWS Glue, and migrated Campaign data from diverse sources to AWS Redshift, achieving a 30% reduction in migration time
* Configured EC2 instances for high availability and security, applied Cloud Watch alerts instances, and Cloud Formation, reducing response time
* Consulting on Snowflake Data Platform Solution Architecture, Design, Development and deployment focused to bring the data driven culture across the enterprises Deployed AES encryption functions within ETL data flows using coding/scripting for column-level transformations
* Provide Business Intelligence support using Tableau and Microsoft Power BI for implementing effective Business dashboards & visualizations of data.
* Designed end to end scalable architecture to solve business problems using various Azure Components like HDInsight, Data Factory, Data Lake, Storage and Machine Learning Studio.
* Used Kubernetes to orchestrate the deployment, scaling and management of Docker containers
* Use Kafka a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for Spark to process data and create Kafka topics for application logs
* Developed a Sqoop script and developed scripts to extract structured data from AWS Redshift, MySQL and other RDS databases data into Hive tables directly on a daily incremental basis.
* Automated OpenStack and AWS deployment using Cloud Formation, Ansible, Chef and Terraform.
* Used terraform to write Infrastructure as code and created Terraform scripts for EC2 instances, Elastic Load balancers and S3 buckets
* Developing Spark applications using Spark - SQL in Databricks for data extraction, Participating immigration Scala code into Microservices
* Executed a serverless architecture using API Gateway, Fargate, Lambda, and Dynamo DB, deployed AWS Lambda code, and heightened processing by 25% by working with SRE teams
* Automated Docker Image building and deployment using AWS Fargate, Continuous Delivery processes, streamlined deployment through Jenkins workflows, and ensured code integrity and quality
* Established data security protocols in AWS through IAM roles and policies to maintain compliance with industry standards and regulations
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC, Parquet, Text Files into AWS Redshift
* Strong experience working with Spark Data frames, Spark SQL and Spark Structured Streaming APIs using Scala
* Developed Snowflake views to load and unload data from and to an AWS S3 bucket, as well as transferring the code to production.
* Strong expertise in using ETL Tool Informatica Power Center 8.6 /9 (Designer, Workflow Manager, Repository Manager, Data Quality (IDQ) and ETL concepts.
* Authored PySpark scripts for Delta table column reordering, modifying data types and detecting incorrectly masked data and enforced Kafka offset checkpointing for resilient job recovery on failures/restarts
* Experience in writing Sub Queries, Stored Procedures, Triggers, Cursors, and Functions on SQL and PostgreSQL database Automated data workflows and scheduling using Databricks notebooks and Databricks Workflows.
* Built robust ETL pipelines to ingest batch and streaming data from diverse sources (Workday, Salesforce, third-party vendor APIs) into Snowflake Data Warehouse handling using Spark, Python, and Snowpipe hosted on AWS EC2.
* Developed real time data streaming solutions using Spark Structured Streaming with Scala applications to consume the JSON messages from Kafka topics.
* Experience in infrastructure configuration and automations process and tools: Terraform, Ansible, Puppet, Chef.
* Analyze machine data and create excel visualization plots as a story narrative for Business Users
* Refactored an event-driven Python Lambda function by connecting enterprise datasets on S3 to the Data Catalog API for automatic schema validation, yielding an 85% increase in data ingestion speed to in-house DataLake.
* Implemented cluster services using Kubernetes and Docker to manage local deployments in Kubernetes by building a self - hosted Kubernetes cluster using Terraform and Ansible and deploying application containers
* Experience in Custom process design of Transformation via Azure Data Factory and Automation Pipelines.
* Extensively used the Azure services Like Azure Data Factory and Logic App for ETL to push in/out data from BLOB Storage, HDFS.
* Automated CI/CD pipelines using GitHub Actions to streamline code deployment and testing processes
* Evaluated and analyzed the architecture and implementation of various AWS services such as Amazon EMR, Step Functions, Redshift, S3, Athena, Glue, and Kinesis experience in end-to-end assessments
* On a day-to-day basis, experienced Agile Scrum methodology and Jira ticket system for project development.

|  |  |
| --- | --- |
| **Data Engineer/Cloud Engineer**  American Express | **Phoenix, AZ**  October 2018 - January 2020 |

* Boosted Spark performance by 40% through optimization of Hadoop algorithms leveraging Spark context, Spark-SQL, Data Frames, pair RDD's, and Spark YARN
* Designed and implemented data pipelines using Google Big Query for large-scale data processing, transformation, and analysis, ensuring efficient and cost-effective data storage and querying
* Implemented data ingestion and event-driven architectures using Google Cloud Storage and Cloud Pub/Sub, enabling seamless integration and processing of data from various sources and systems
* Configured Kafka topics across environments, enhancing data streaming by processing JSON messages per second
* Developed test scripts for continuous integration, ensuring code quality and functionality, and improving test coverage.
* Created Teradata schemas with constraints, Created Macros in Teradata. Loaded the data using Fast load utility. Created functions and procedures in Teradata.
* Engaged with various Big Data components such as Spring, Hive, Apache Spark, Tableau Postgres SQL, SQOOP, HBase, Kafka, and Docker for efficient data processing and storage
* Designed and built pipelines, facilitating data analysis with an emphasis on efficiency and improving test coverage
* Implemented Kubernetes solutions to optimize resources, reducing infrastructure costs by 25% with high performance
* Proficient in utilizing Snowflake features such as SnowSQL and SNOWPIPE for seamless and continuous data ingestion for analytical purposes
* Strong experience working with Spark Data frames, Spark SQL and Spark Structured Streaming APIs using Scala
* Implemented batch processing of jobs using Spark Scala API's
* Developed real time data streaming solutions using Spark Structured Streaming with Scala applications to consume the JSON messages from Kafka topics
* Used rest API with Python to ingest Data from and some other site to BIGQUERY
* Monitoring Big query, Data proc and cloud Data flow jobs via Stack driver for all the environments
* Experience in handling the applications authentications by using Oauth2/SSO
* Utilized Python's powerful libraries like NumPy, Pandas, and sci-kit-learn to pre-process and pre-process and analyze, querying large datasets effectively
* Worked on creating data ingestion processes to maintain Data Lake on the GCP cloud and Big Query
* Conducted troubleshooting and debugging of issues related to Snowflake data warehouse, identifying root causes and implementing solutions to minimize downtime and optimize system performance
* Experience in defining, designing, and developing Java applications, specially by leveraging frameworks such as Spark, Kafka
* Improved the performance of the Kafka cluster by fine tuning the Kafka Configurations at producer, consumer and broker level
* Worked on Agile Project methodology and was involved in activities like Program Increments, Story discussions, Sprint planning, and daily scrums.

|  |  |
| --- | --- |
| **Big Data Engineer- Hadoop Developer**  Chemical Abstract Services | **Columbus, OH**  October 2016 - October 2018 |

* Orchestrated and automated data engineering workflows using Google Cloud Composer by Apache Airflow, enabling workflow scheduling, monitoring, and orchestration for data processing and ETL/ELT tasks
* Integrated Hive and HBase, executing MR Unit tests and improving MapReduce job accuracy
* Led the transformation of Mainframe tables to HDFS and HBase using Sqoop and Pentaho kettle, handling 5TB data.
* Developed spark applications in data bricks using pyspark and spark SQL to perform transformations and aggregation on source data before loading it into Azure Synapse Analytics for reporting
* Working Experience on Azure Data bricks cloud to organize the data into notebooks and making it easy to visualize data using dashboards
* Knowledge of Azure cloud services storage, Data factory, data lake, Logic Apps, Azure monitoring, Active directory, synapse, key vault and SQL Azure.
* Build the data pipeline using Azure services like Data Factory to load data from SQL Server to Azure Data Base using Data Factories, API Gateway Services and Python codes.
* Revamped application logic in Azure Data Lake and Data Factory, migrating functionalities to SQL Database and SQL data warehouse environments
* Experienced with CAS's different domains data which were build using the priority order of XSLT, XQuery
* Designed and deployed data pipelines and workflows for extracting, transforming, and loading (ETL) medical and pharmacy claims data into data warehouses or data lakes
* Conducted Spark ETL pipeline delivery on Azure Databricks, orchestrating data transformation via Azure Data Factory, achieving a good improvement in data processing
* Subsidized to at-scale machine learning systems, facilitating predictive analysis with a 20% improvement in accuracy
* Collaborated on GCP, utilizing Hadoop, Pig, Flume, and Oozie for data management in a distributed system
* Advanced and integrated Web Services using Java/J2EE in WebLogic, enhancing application performance
* Leveraged Python on Linux and UNIX platforms, developing robust web applications for efficient data processing
* Crafted web pages using programming languages Java Servlet, JSP, CSS, JavaScript, DHTML, and HTML incorporating Struts validation for improved user experience
* Experience with Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD's, Spark YARN, Ansible, Terraform, GIT, Chef
* Automated resulting scripts and workflow using Apache Airflow
* Managing messages on Kafka topics using Talend.
* Mainly worked on Big Data components like Spring boot, Hive, Apache Spark, Tableau, Postgres SQL, SQOOP & HBase, Kafka, and Docker for data processing and storage
* Responsible for testing & debugging code to enhance performance by using the latest Libraries through Spark and Scala
* Utilized Jinja templating language within dbt to dynamically generate SQL queries and configurations based on parameters and variables, enabling parameterization and customization of data transformation processes
* Experience on log parsing, complex Splunk searches, including external table lookups, Splunk data flow, components, features and product capability
* Implemented data modeling, partitioning, and indexing strategies in Azure Synapse Analytics to optimize query performance and improve data processing efficiency
* Hands on experience in developing SPARK applications using Spark tools like RDD transformations, Spark core, Spark MLlib, Spark Streaming and Spark SQL
* Used AWS ECS to run docker containers on ECS Fargate, and ECS on EC2 and implemented docker container logging with docker sidecar container
* Managed and analyzed medical and pharmacy claims data to identify trends, optimize reimbursement processes, and support data-driven healthcare decisions
* Good experience in creating and designing data ingest pipelines using technologies such as Apache Storm- Kafka
* Performed Data Analysis using SQL, PL/SQL, Python, Spark, Databricks, Teradata SQL Assistant, SQL server management studio, SAS
* Used Python programming and Django for the backend development, Bootstrap and Angular for frontend connectivity and MongoDB for database
* worked extensively with automation tools like Jenkins, Artifactory, SonarQube to implement End-to-End Automation
* Played an active role in code reviews, promoting code quality and teamwork
* Adept at problem-solving and communication, actively collaborating with teams to initiate effective solutions
* Good experience working with ON-Prem Clusters like Horton Works and Cloudera distribution
* Experience in all stages of SDLC (Agile, Waterfall), writing Technical Design documents, Testing, Debugging, and Implementation of enterprise-level Datamart & Data Lake house
* Worked closely with data scientists to build and deploy machine learning models using Databricks MLflow
* Worked extensively with Dimensional modeling, Data migration, Data cleansing, Data Profiling Analytics, and ETL Process features for Data Warehouses.

|  |  |
| --- | --- |
| **SoftWare Engineer**  Toll Plus Inc. | **Hyderabad, IN**  April 2014 - December 2015 |

* Designed and developed Web Services using Java/J2EE in WebLogic environment. Developed web pages using Java Servlet, JSP, CSS, Java Script, DHTML, and HTML. Added extensive Struts validation. Wrote Ant scripts to build and deploy the application.
* Involve in the Analysis, Design, and Development and Unit testing of business requirements.
* Developed business logic in JAVA/J2EE technology.
* Implemented business logic and generated WSDL for those web services using SOAP.
* Worked on Developing JSP pages, Implemented Struts Framework.
* Modified Stored Procedures in Oracle Database.
* Developed the application using Spring Web MVC framework.
* Worked on the Spring DAO module and ORM using Hibernate. Used Hibernate Template and Hibernate Dao Support for Spring-Hibernate Communication.
* Configured Association Mappings such as one-one and one-many in Hibernate
* Worked with JavaScript calls as the Search is triggered through JS calls when a Search key is entered in the Search window
* Worked on XML, XSL and XHTML files.

Education

|  |  |
| --- | --- |
| **Master of Science in Computer Technology**  Eastern Illinois University, **IL** |  |

|  |  |
| --- | --- |
| **Bachelors in computer science & Engineering**  J.N.T. University |  |

Certifications

|  |  |
| --- | --- |
| * AWS Certified Solutions Architect – Associate | * Azure Developer associate |
| * Python Certification | * Data bricks Analysis Certification |

Website, Portfolio and Profiles

https://www.linkedin.com/in/vijay-kumar-6807a2162/

APPLICATION RESPONSIBILITIES:

* Exploring with Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD's, Spark YARN, Ansible, Terraform, GIT, Chef.
* Automated resulting scripts and workflow using Apache Airflow. Managing messages on Kafka topics using Talend.
* Mainly worked on Big Data components like Spring boot, Hive, Apache Spark, Tableau, Postgres SQL, SQOOP & HBase, Kafka, and Docker for data processing and storage.
* Responsible for testing & debugging code to enhance performance by using the latest Libraries through Spark and Scala.
* Utilized Jinja templating language within dbt to dynamically generate SQL queries and configurations based on parameters and variables, enabling parameterization and customization of data transformation processes.
* Experience on log parsing, complex Splunk searches, including external table lookups, Splunk data flow, components, features and product capability.
* Implemented data modeling, partitioning, and indexing strategies in Azure Synapse Analytics to optimize query performance and improve data processing efficiency.
* Hands on experience in developing SPARK applications using Spark tools like RDD transformations, Spark core, Spark MLlib, Spark Streaming and Spark SQL.
* Used AWS ECS to run docker containers on ECS Fargate, and ECS on EC2 and implemented docker container logging with docker sidecar container
* Managed and analyzed medical and pharmacy claims data to identify trends, optimize reimbursement processes, and support data-driven healthcare decisions.
* Good experience in creating and designing data ingest pipelines using technologies such as Apache Storm- Kafka.
* Performed Data Analysis using SQL, PL/SQL, Python, Spark, Databricks, Teradata SQL Assistant, SQL server management studio, SAS.
* Used Python programming and Django for the backend development, Bootstrap and Angular for frontend connectivity and MongoDB for database.
* Worked extensively with automation tools like Jenkins, Artifactory, SonarQube to implement End-to-End Automation.
* Played an active role in code reviews, promoting code quality and teamwork.
* Adept at problem-solving and communication, actively collaborating with teams to initiate effective solutions
* Good experience working with ON-Prem Clusters like Horton Works and Cloudera distribution.
* Experience in all stages of SDLC (Agile, Waterfall), writing Technical Design documents, Testing, Debugging, and Implementation of enterprise-level Datamart & Data Lake house.
* Implemented large Lambda architectures using Azure Data platform capabilities like Azure Data Lake, Azure Data Factory, HDInsight, Azure SQL Server, Azure ML and Power BI.
* Worked extensively with Dimensional modeling, Data migration, Data cleansing, Data Profiling Analytics, and ETL Process features for Data Warehouses.
* Excellent analytical and communication skills with the capability to handle new technologies.
* Extensive Knowledge of Google Cloud Platform (GCP) services like compute engine, cloud load balancing cloud storage, cloud Data proc, Cloud Pub/Sub, SQL BigQuery, stack driver monitoring, Cloud spanner, looker, deployment manager.
* Worked closely with data scientists to build and deploy machine learning models using Databricks MLflow.
* Developed Json Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
* Hands-on experience in Graph Database language Relational AI technology. (GQL).
* Experience in working on NoSQL databases – Hbase, Cassandra and MongoDB, database performance tuning and data modeling.
* Optimized SnowSQL scripts with clustering, CTEs, and indexing for data quality checks, reducing duplicate data by 63%.
* Extensively used the Azure services Like Azure Data Factory and Logic App for ETL to push in/out data from BLOB Storage, HDFS
* Provide Business Intelligence support, and Tableau for implementing effective Business dashboards visualizations of data.
* Extensive experience in using Microsoft BI studio products like SSIS, SSAS, and SSRS, Airbyte for implementation of ETL methodology in data extraction, transformation, and loading.
* Implemented scalable and cost-effective data storage solutions using Azure Blob Storage and Azure Data Lake Storage (ADLS), enabling reliable and secure storage of structured and unstructured data.
* Experience in AWS platform features including IAM, EC2, EBS, VPC, Step Functions, RDS, S3, SQS, Redshift, Lambda, AWS, Fargate CloudWatch, Cloud Trail, CloudFormation AWS Configuration, Autoscaling, EMR, SNS, Route53, Containers and Kubernetes.