**Vinay Kumar**

**Sr. Bigdata AWS Cloud Engineer**

**Email:** [saisri@mastroservices.com](mailto:saisri@mastroservices.com)

**Phone: +1 (469 846 8345)**

**SUMMARY:**

* 9+ years of experience in IT which includes Analysis, Design, Data Mapping, Datastage, Development of Big Data using Hadoop, AWS, Python, data Lake, Scala, GCP, design and development of web applications using database, and data warehousing development using My SQL, Oracle.
* Experience in creating applications using Spark with Python.
* Experienced in Apache Spark and developing data processing and analysis algorithms using Python.
* Strong working experience in Big Data Analytics with hands-on experience in installing, configuring, and using ecosystem components like Hadoop Map Reduce, HDFS, HBase, Zookeeper, Hive, Sqoop, Pig, Flume, Cassandra, Kafka, and Spark.
* Building data processing triggers for Amazon S3 using AWS Lambda functions with Python.
* Worked on the Spark SQL and Spark Streaming modules of Spark and used Scala and Python to write code for all Spark use cases.
* Strong knowledge of Data Extraction, Data Mapping, Data Conversion, and Data Loading processes using UNIX Shell scripting, SQL, PL/SQL, and SQL Loader.
* Experience in Migrating SQL database to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse and controlling and granting database access and Migrating On-premise databases to Azure Data Lake store using Azure Data Factory.
* Good Understanding of Hadoop architecture and Hands-on experience with Hadoop components such as Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce concepts, and HDFS Framework.
* Expertise in preparing Interactive Data Visualization using Tableau software from different sources.
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* Implemented advanced procedures like text analytics and processing using in-memory computing capabilities like Apache Spark written in Scala.
* Extensively worked on AWS services like EC2, S3, EMR, Lambda, Cloud watch, RDS, Auto scaling, Cloud Formation, SQS, ECS, EFS, DynamoDB, Route53, Glue etc.
* Extensive experience working with AWS services such as Amazon Redshift, Amazon RDS (equivalent to Cloud SQL), Amazon EMR (equivalent to Dataproc), Amazon S3, and Amazon DynamoDB.
* Proficient in AWS Big Data Services, including Redshift, EMR, and S3, for enterprise-level data modeling and implementation, ensuring efficient data management.
* Designed and developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Demonstrated expertise in setting up Python REST API Framework using Django for efficient web service development.
* Utilized Stackdriver service and Dataproc clusters in GCP for accessing logs for debugging purposes, enhancing operational efficiency.
* Proficient in a wide range of AWS services including VPC, Glue Pipelines, Glue Crawler, CloudFront, EC2, ECS, EKS, Elastic Beanstalk, Lambda, S3 Storage, RDS, DynamoDB, Redshift, Elastic Cache, DMS, SMS, Data Pipeline, IAM, WAF, Artifacts, API Gateway, SNS, SQS, SES, Auto Scaling, CloudFormation, CloudWatch, and CloudTrail
* Managed operations and maintenance support for AWS cloud resources, including launching, maintaining, and troubleshooting EC2 instances, S3 buckets, Auto Scaling, DynamoDB, AWS IAM, Elastic Load Balancers (ELB), and Relational Database Services (RDS). Additionally, created snapshots for data storage in AWS S3.
* Experience in Data load management, importing & exporting data using SQOOP&FLUME.
* Experience in dealing with log files to extract data and copy into HDFS using Flume.
* Experience in integrating Hive and Hbase for effective operations.
* Experience in Impala, Solr, MongoDB, HBase, Spark, and Kubernetes.
* Expertise in Waterfall and Agile-SCRUM methodologies.
* Experienced with code versioning and dependency management systems such as Git, SVT, and Maven.
* Experience in Testing and documenting software for client applications.
* Good experience in using Data Modelling techniques to find the results based on SQL and PL/SQL queries.
* Experience working with different databases, such as Oracle, SQL Server, and MySQL, and writing stored procedures, functions, joins, and triggers for different Data Models.
* Experienced in handling different file formats like Text files, Avro data files, Sequence files, XML, and JSON files.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Cloud Technologies | Azure, AWS |
| Azure Skills | Azure Data Lake, Data Factory, Databricks, Azure SQL Databases, Azure SQL Datawarehouse |
| AWS Skills | S3, EC2, EMR, Glue, Athena, Aurora, SNS, Crawler |
| Big data Skills | HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, HBase, Spark, Spark Streaming, Yarn, Zookeeper, Kafka, ETL |
| Programming Languages | Python, Spark, R, Scala, Terraform, Angular |
| Databases | Aurora DB, Oracle, MySQL |
| Job Scheduler Tools | Airflow |
| CICD | Bit Bucket, Jenkins |
| IDE | PyCharm, Visual Studio |

**PROFESSIONAL EXPERIENCE:**

***Role: Sr Bigdata AWS Cloud Engineer February 2023 – Present***

***Client: CapitalOne, Chicago, IL***

**Responsibilities:**

* Created various data pipelines using Spark, Scala, and SparkSQL for faster processing of data
* Created data mappings based on the business requirements and translated them into technical solutions.
* Performed Analysis on Data mapping between different data models.
* Designed number of partitions and replication factors for Kafka topics based on business requirements and worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using Python (PySpark).
* Created Data Quality Scripts using SQL and Hive to validate the successful das ta load and quality of the data. Created various types of data visualizations using Python and Tableau.
* Extensively worked with Avro and Parquet, XML, and JSON files and converted the data from either format Parsed Semi-Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Involved in working with Spark on top of Yarn/MRv2 for interactive and Batch Analysis
* Worked with a team to migrate from Legacy/On-prem environment into AWS.
* Stored and retrieved data from data warehouses using Amazon Redshift.
* Experienced in analyzing and Optimizing RDDs by controlling partitions for the given data
* Experienced in writing live Real-time Processing using Spark Streaming with Kafka
* Used HiveQL to analyze the partitioned and bucketed data and compute various metrics for reporting
* Worked with querying data using SparkSQL on top of the Spark engine.
* Involved in managing and monitoring the Hadoop cluster using Cloudera Manager.
* Used Python and Shell scripting to build pipelines.
* Developed data pipeline using SQOOP, HQL, Spark, and Kafka to ingest Enterprise message delivery data into HDFS.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Assisted in creating and maintaining technical documentation to launching HADOOP Clusters and even for executing Hive queries and Pig Scripts.
* Assisted in Cluster maintenance, cluster monitoring, adding and removing cluster nodes and installed and configured Hadoop, Map Reduce, HDFS, and developed multiple Map Reduce jobs in SQL for data cleaning and pre-processing.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Created S3 buckets (configure, policies, permissions) and used AWS S3 for storage and backup of Data into AWS and AWS Glacier to store archive data.
* Implemented AWS solutions using E2C, S3, RDS, EBS, Elastic Load Balancer, Glue Pipelines, Glue Crawler, Auto scaling groups, Optimized volumes, and EC2 instances and created monitors, alarms, and notifications for EC2 hosts using Cloud Watch.
* Implemented AWS Step Functions to automate and orchestrate the Amazon PageMaker-related tasks such as publishing data to S3, training the ML model, and deploying it for prediction.
* Handled AWS EC2, Amazon S3, Amazon RDS, Elastic Load Balancer, Auto Scaling, CloudWatch, SNS, Amazon Lambda andStepFunctions.
* Generated consumer group lags from Kafka using their API Kafka. Used for building real-time data pipelines between clusters. Extracted files from Hadoop and dropped them on a daily hourly basis into S3.
* Automated and monitored complete AWS infrastructure with Terraform.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.

***Role: Bigdata Engineer February 2019 – December 2021***

***Employer: OathMap LLP, Hyderabad, India***

***Client: XILINX (Contractor)***

**Responsibilities:**

* Proficient in multi-cloud architecture, designing and building end-to-end ETL pipelines in both GCP and AWS environments, encompassing data ingestion and transformation.
* Implemented BigQuery data processing from GCP Pub/Sub topics to BigQuery using Cloud Dataflow with Python, and utilized REST API with Python for data ingestion from various systems into BigQuery.
* Executed Cloud Dataflow for data validation between raw source files and BigQuery, integrated monitoring of BigQuery and Dataproc using Stackdriver across environments.
* Developed Perl scripts for data feed handling, implementing business logic, and communicating with web services through SOAPLite module and WSDL.
* Generated data cubes using Hive, Pig, and Java MapReduce on provisioned Hadoop clusters in AWS.
* Created S3 buckets, configured policies, and permissions, utilizing AWS S3 for data storage and backup, and AWS Glacier for archiving data.
* Implemented various AWS solutions including EC2, S3, RDS, EBS, Elastic Load Balancer, Glue Pipelines, Glue Crawler, Auto Scaling groups, and EC2 instances, along with setting up monitors, alarms, and notifications using CloudWatch.
* Utilized AWS Step Functions to automate and orchestrate Amazon Sagemaker tasks such as publishing data to S3, training ML models, and deploying them for prediction.
* Managed AWS services including EC2, S3, RDS, Elastic Load Balancer, Auto Scaling, CloudWatch, SNS, Amazon Lambda, and Step Functions.
* Generated consumer group lags from Kafka API for building real-time data pipelines between clusters and extracted files from Hadoop into S3 on a daily hourly basis.
* Authored Python (PySpark) scripts for custom UDFs, row/column manipulations, merges, aggregations, stacking, data labeling, and cleaning tasks.
* Migrated an entire Oracle database to BigQuery and built data pipelines in Airflow for ETL jobs using different Airflow operators.
* Proficient in data modeling including dimensional data modeling, Star Schema/Snowflake schema, FACT & Dimension tables, physical & logical data modeling, utilizing tools like Data Model and Erwin.
* Developed Kibana Dashboards based on Logstash data, and integrated various source and target systems into Elasticsearch for near real-time log analysis of end-to-end transactions.
* Worked with Informatica PowerCenter for mappings, mapping parameters, workflows, variables, and session parameters.
* Wrote Pig Scripts to generate MapReduce jobs and performed ETL procedures on data in HDFS.
* Developed solutions leveraging ETL tools and identified opportunities for process improvements using Informatica and Python.
* Actively participated in Scrum teams and Agile practices including Test-Driven Development (TDD), Behavior-Driven Development (BDD), and pair programming.
* Proficient in automation tools like GIT and dimensional modeling including Star Schema, Snowflake schema, and transactional modeling.
* Conducted reviews of existing systems and provided unified feedback on jobs, along with testing databases using complex SQL scripts and effectively handling performance issues.

***Role: Bigdata Engineer December 2017 – February 2019***

***Employer: OathMap LLP, Hyderabad, India***

***Client: AT&T (Contractor)***

**Responsibilities:**

* Designed and implemented end-to-end data pipelines in Azure Data Factory (ADF), leveraging Linked Services, Datasets, and Pipelines for seamless Extract, Transform, and Load (ETL) processes across diverse data sources including Azure SQL, and Azure SQL Data Warehouse.
* Orchestrated real-time streaming pipelines from DB2 to HDFS utilizing Apache Kafka, ensuring efficient data ingestion and processing.
* Practiced Agile Scrum methodology, facilitating iterative application development through weekly Sprints and contributing to stand-up meetings for streamlined project progress tracking.
* Proficient in creating and managing both External and Managed Hive tables, utilizing HiveQL for data querying and manipulation within Hadoop ecosystems.
* MapReduceed comprehensive analysis and documentation of both new and existing systems to ensure clarity and maintainability of data processes.
* Validated MapReduce, Pig, and Hive Scripts by extracting data from Hadoop, comparing against file and report data for accuracy, ensuring data integrity and equal access.
* Developed a new Web Service and Client using Spring-WS to access alternate contractor details, enhancing system functionality and accessibility.
* Utilized JIRA for meticulous issue tracking and bug reporting, ensuring swift resolution and efficient project management.
* Implemented Spark using Scala and Spark SQL for expedited data processing, demonstrating expertise in managing and analyzing data from various sources.
* Engineered data pipelines and ETL workflows within Azure Synapse Analytics, leveraging Azure Data Factory, Azure Databricks, and Azure Synapse Studio for seamless data integration and deployment.
* Built and deployed Azure Synapse Analytics pipelines and workflows using Azure DevOps and Azure Resource Manager templates, ensuring efficient deployment and management.
* Integrated Azure Synapse Analytics with other Azure services such as Azure Blob Storage, Azure Event Hubs, and Azure Key Vault, enabling seamless data exchange and security.
* Established secure connections between Hadoop clusters and Azure Data Factory by implementing SELF-HOSTED integration runtime on Windows servers, facilitating smooth data migration from HDFS to Azure Data Lake.
* Implemented versioning through Azure GIT on Azure Data Factory and scheduled pipelines through Scheduled Triggers, ensuring robust version control and timely execution of workflows.
* Collaborated effectively with the team using GIT, GitHub, and SourceTree version control platform, ensuring seamless code management and collaboration.
* Developed RESTful APIs using Bottle micro-framework with MongoDB (NoSQL) as the backend database, enabling efficient data retrieval and manipulation.
* Developed Spark scripts using Scala shell commands and Spark SQL for optimized data processing and testing, contributing to enhanced performance and reliability of data workflows.

***Role: Hadoop Developer January 2016 – November 2017***

***Employer: OathMap LLP, Hyderabad, India***

***Client: Headstrong (Contractor)***

**Responsibilities:**

* Developed MapReduce programs for filtering and preprocessing unstructured data, enhancing data quality and usability.
* Proficient in handling Hive partitions and buckets for optimized data storage and retrieval.
* Experienced in managing Hive schema evolution with Avro file format to ensure data compatibility over time.
* Skilled in processing semi-structured/serialized data using Hive (Avro, Parquet, ORC) for diverse data sources.
* Expertise in creating and managing various Hive table types (managed, external, partitioned) to support business requirements.
* Designed efficient data models with Hive tables and partitions, optimizing query performance
* Experienced in importing/exporting large datasets between Hadoop and relational databases using Sqoop.
* Adept in scheduling and automating Sqoop jobs for incremental runs, streamlining workflow processes.
* Developed automated data pipelines using Sqoop, Oozie, and Hadoop ecosystem components, enhancing scalability and efficiency.
* Optimized PL/SQL queries to improve the performance of stored procedures, enhancing overall system efficiency.
* Exported data from HDFS to RDBMS using Sqoop for Business Intelligence, visualization, and reporting purposes.
* Managed Hadoop clusters using Cloudera Manager, ensuring smooth operation and efficient resource utilization.
* Translated HiveQL into Spark transformations using Spark RDD and Scala programming, optimizing data processing workflows.
* Integrated Kafka-Spark streaming for high-throughput and reliable data processing, enhancing system efficiency.
* Tuned Hive and Pig to improve performance and resolve performance issues, optimizing query execution and data processing.

***Role: ETL Developer September 2013- December 2015***

***Client: Squires Solutions, Hyderabad, India***

**Responsibilities:**

* Analyzed project requirements and conducted Impact Analysis to determine the effect on existing ETL processes.
* Developed DDL and DML scripts to transform source data and load it into target tables, ensuring data integrity and consistency.
* Implemented data transformations using SQL queries based on business logic outlined in the Mapping sheet.
* Modified and optimized existing stored procedures and ETL workflows in Microsoft SQL Server Management Studio and Control-M scheduler to accommodate new business requirements.
* Conducted thorough peer reviews of SQL scripts, Control-M jobs, and Informatica workflows to ensure adherence to coding standards and best practices.
* Collaborated with the QA team during testing phases, assisting in identifying and resolving defects to ensure the quality and reliability of ETL processes.
* Designed and developed Informatica Workflow to extract data from XML files and load it into the database.
* Improved session Performance by enabling property incremental aggregation to load incremental data into the target table.
* Involved in the continuous enhancements and fixing of production problems. Designed, Implemented, and Tuned interfaces and batch jobs using PL/SQL.
* Involved in data replication and high availability design scenarios with Oracle Streams. Developed UNIX Shell Scripts to automate repetitive database processes.

**EDUCATION QUALIFICATIONS:**

* Master’s in computer technology From EIU-2023.