**Deekshita Gundrati**

Sr. Data Engineer

Email: deekshita238@gmail.com

Pradeep.k@tekprosol.com

Phone:  609-739-8753

**PROFESSIONAL SUMMARY**

* 8+ years experienced Analytic Data Engineer with a demonstrated professional working Experience in Big Data Technologies like Ingestion, Data Modelling, Querying, Processing, Analysis and Implementing Enterprise level Systems Spanning Big Data and Data Integration
* Hands on Experience on Hadoop Distribution Platforms Namely IBM Big Insights, Hortonworks and Cloudera and Cloud platforms GCP and AWS
* Expertise in Big Data Technologies and Hadoop Ecosystems such as Pyspark, Spark - Scala, HDFS, GPFS, Hive, Sqoop, PIG, Spark-SQL, Kafka, Hue, Yarn, Trifacta and EPIC data sources.
* Good knowledge on Amazon AWS concepts like EMR and EC2 web services which provides fast and efficient processing of Big Data and Machine Learning Concepts.
* Hands on experience in Building Data pipelines and Data marts using Hadoop stack.
* Hands on experience in Apache Spark creating RDD’s and Data Frames applying Operations Transformation and Actions and concerting RDD’s to Data Frames.
* Experienced in data processing like collecting, aggregating, moving from various sources using Apache Flume and Kafka.
* Implementing scalable data processing and analytics solutions using Apache Spark, Kafka, and Java
* Experience in writing REST APIs in Python for large-scale applications.
* Extensive experience working with AWS Cloud services and AWS SDKs to work with services like AWS API Gateway, Lambda, S3, IAM and EC2.
* Developed a data pipeline using Kafka and Spark Streaming to store data into HDFS and performed the real-time analytics on the incoming data.
* Designed, developed, and implemented a real-time data pipeline using Apache Spark Streaming and Apache Kafka to process and analyze real-time customer data.
* In depth understanding of Apache Spark job execution components like DAG, Executors, Task Schedular, Stages and Spark Stearming.
* Experience in Creating and executing Data Pipelines in GCP and AWS platforms.
* Hands on Experinece in GCP, Big query, cloud functions, data proc.
* Strong Experience in Control-M Job Scheduler Tool, Apache Air flow, ESP, D-series. Monitored the jobs on call base to close Incident tickets.
* Hands-on experience withAmazon EC2, S3, RDS, IAM, Auto Scaling, CloudWatch, SNS, Athena, Glue, Kinesis, Lambda, EMR, Redshift, DynamoDB and other services of the AWS family.
* Expertise in using CICID JENKINS pipeline to deploy the codes into production
* Designed and developed the program paradigm to support data collection, filtering process in data warehouse and Hadoop data mart.
* Strong Hadoop and platform support experience with all the entire suite of tools and services in majorHadoop Distributions- Cloudera, Amazon EMR, and Hortonworks.
* Hands-on Experience in working with globally distributed team in Europe, Mexico & India and AGILE Implementation methodology.
* Deep understanding of Cyber security, pen testing and working with them to get approvals to deploy code into production.
* Hands on experience in working Agile environment and follow release management, Golden rules.
* Experience in Version control tools such as GIT and Urban Code Deployment (UCD) tools.

**Skills**:

|  |
| --- |
| **Programming Languages:**PYTHON, Scala, SHELL SCRIPTING, Java |
| **Big Data Eco-System:**HDFS, GPFS, Hive, Sqoop, Spark, Yark, PIG, Kafka |
| **Operating Systems:**Windows, Linux (Centos, Ubuntu) |
| **Hadoop Distributions:**Hortonworks, Cloudera, IBM Big Insights |
| **Databases**Hive, MYSQL, NETEZZA, SQL Server |
| **IDE Tools & Utilities:**IntelliJ IDEA, Eclipse, PyCharm, Aginity Workbench, GIT |
| **Markup Languages:**HTML |
| **Job Scheduler:**Control-M, IBM Symphony Platform, Ambari, Apache Air flow |
| **Cloud Computing Tools:**GCP, AWS, Snowflake. |
| **Scrum Methodologies:**Agile, Asana, Jira |
| **Others:**MS Office, RTC, Service Now, OPTIM, IGC (Info sphere Governance catalog), WinSCP, MS Visio |

**Professional Experience:**

**Role: Senior Big Data Engineer Oct 2021 - Present**

**Company: Lincoln Financial Group Greensboro, NC**

**Responsibilities:**

* Building Data Stream Lines in Google Cloud Platform (Iot Registry, Pub/Sub, DataFlow, BigQuery, DataPrep, Data Studio, AI Platform)
* Developed applications and deployed them in Google Cloud Platform using DataProc, Dataflow, Composer, BigQuery, BigTable, Cloud Storage, GCS and various operators in DAG.
* Migrated existing data pipelines in hive to GCP platform
* Designed and implemented data transformation, ingestion and curation functions on GCP cloud using GCP native and Python.
* Optimized data pipelines for performance and cost for large scale data lakes.
* Designed and automated Big Query tables and Google Cloud Functions to enable reporting, analysis, and modeling.
* Used Node.js to write custom UDFs in Big query and used them in the data pipeline.
* Used Python for scripting purposes, for leveraging a wide range of technologies that include leveraging a wide range of technologies
* Worked on Developing and supporting databases and related ETL (batch and real-time processing)
* Good understanding of issue triaging and resolution protocols in Big Data systems
* Designing, Testing and Implementing data migration/ingestion/processing/quality frameworks that will be able to handle hundreds of GBs data using Airflow, Pyspark, PythonandBigquery.
* Conduct design and code reviews to ensure high quality of work is delivered.
* Constantly engaging with data customers to get feedback around the data solutions developed and building documentation for the data engineering best practices.
* Developed Hive Scripts, Hive UDFs, Python Scripting and used Spark (Spark-SQL, Spark-shell) to process data in Hortonworks.
* Developer and maintained java based ETL processes, improving data extraction efficiently by 20%
* Built a system for analyzing the column names from all tables and identifying personal information columns of data across on-premises Databases (data migration) to GCP
* Designed and Developed Scala code for data pull from cloud-based systems and applying transformations on it.
* Usage of Sqoop to import data into HDFS from MySQL database and vice-versa.
* Implemented optimized joins to perform analysis on different data sets using MapReduce programs.
* Created continuous integration and continuous delivery (CI/CD) pipeline on AWS that helps to automate steps in software delivery process.
* Experience in processing of load and transform the large data sets of structured, unstructured and semi structured data in Hortonworks.
* Implemented Partitioning, Dynamic Partitions and Buckets in HIVE & Impala for efficient data access.
* Worked inAgile environment and used rally tool to maintain the user stories and tasks.
* Extensively worked on HiveQL, join operations, writing custom UDF's and having good experience in optimizing Hive Queries.
* Performed analysis on data discrepancies and recommended solutions based upon root cause.
* Designed and developed job flow using Apache Air flow.

Environment: HDFS, Python Scripting, Map Reduce, Hive, Impala, Spark-SQL, Spark Streaming, Sqoop, AWS S3, Java, GCP, BigQuery, JDBC, AWS, Python, Scala, UNIX Shell Scripting, Git.

**Role: Data Engineer**

**Company: Honeywell Chicago, IL Mar 2018 - Sep 2021**

**Responsibilities:**

* Developed PySpark pipelines which transforms the raw data from several formats to parquet files for consumption by downstream systems.
* Assisted in the development of data pipelines using Java and Python for
* Used AWS Glue services like crawlers and ETL jobs to catalog all the parquet files and make transformations over data according to the business needs.
* Worked with AWS services like S3, Glue, EMR, SNS, SQS, Lambda, EC2, RDS and Athena to process data for the downstream customers.
* Created libraries and SDKs which will be helpful in making JDBC connections to hive database and query the data using Play framework and various AWS services.
* Developed scripts using Spark which are used to load the data from Hive to Amazon RDS(Aurora) at a faster rate.
* Created views on top of data in Hive which will be used by the application using Spark SQL.
* Applied security on data using Apache Ranger to set row level filters and group level policies on data.
* Normalized the data according to the business needs like data cleansing, modifying the datatypes and various transformations using Spark, Scala and AWS EMR.
* Worked on creating the CI/CD pipelines using tools like Jenkins and Rundeck which will be responsible for scheduling the daily jobs.
* Developed Sqoop jobs which will be responsible for importing the data from Oracle to AWS S3.
* Developed a utility which transforms and exports the data from AWS S3 to AWS glue and sends alerts and notifications to downstream systems (AI and Data Analytics) once the data is ready for usage.
* Worked on groovy scripted Jenkins CICD pipelines, to automate Hadoop cluster scaling. Provisioned servers and deployed features using Ansible playbooks.
* Worked on Jenkins for CICD, pull code from version controls like GitHub, built apache maven and Gradle. Built artifacts are stored in repositories like nexus.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Developed pipelines for auditing the metrics of all applications using AWS Lambda, Kinesis Firehoses.
* Developed end to end pipeline which exports the data from parquet files in S3 to Amazon RDS.
* Worked on optimizing performance of Hive queries using Hive LLAP and various other techniques.

Environment: AWS, Spark, Pyspark, Python, Hadoop, Hive, Sqoop, Play framework, Apache Ranger, S3, EMR, EC2, SNS, SQS, Lambda, Zeppelin, Kinesis, Athena, Jenkins CICD, Rundeck and AWS Glue.

**Role: Data Engineer**   **Oct 2016 – Dec 2017**

**Company: Pactera Edge Chicago, IL**

**Responsibilities:**

* Adept in Agile project management methodology and SDLC (Software Development Life cycle) Requirement gathering, analysis, Design, Development and testing of application using AGILE and SCRUM methodology.
* Detailed understanding on existing build system and tools related to information of various products, release and test results.
* Expertise in process improvement, data extraction, data cleansing, SCRUM data manipulation, Normalization and Denormalization concepts and principles.
* Created ETL data pipelines in state-of-the-art AWS environment EC2, S3, Lambda, etc with AWS Glue.
* Worked on Amazon Web Services (EC2, ELB, VPC, S3, CloudFront, Elasticsearch, IAM, RDS, Route 53, CloudWatch, SNS, Redshift, kinesis, RDS, Lambda, Glue, SageMaker, Personalize).
* Setting up and configuring AWS Virtual Private Cloud (VPC) Components—subnets, IGW, Security Groups. EC2 Instances, Elastic Load Balancers & NAT Gateways for an Elastic Map Reduce Cluster.
* Deploying, managing, and operating scalable, highly available, and fault tolerant systems on AWS
* Actively managed the day-to-day AWS accounts, make recommendations on how best to support our global infrastructure and interact with Developers and Architects in cross functional areas
* Hands on experience with AWS CLI interface and designing Scalable AWS solutions.
* Strong Experience in implementing Data warehouse solutions in Redshift; Worked on various projects to migrate data from on premise databases to Redshift, RDS and S3.
* Hands on experience working knowledge on AWS SageMaker.
* Developed and deployed a product recommendation system using AWS SageMaker based on Matrix Factorization and KNN algorithms.
* Researched extensively on AWS Personalize and deployed the event ingestion code snippet on our product website.
* Assisted in implementation and maintenance of security and data encryption technologies.
* Conducted complete analysis of database capacity and performance requirements.
* Experience building reusable ETL components using Postgres and snowflake.
* Worked extensively on writing triggering Snowpipe, Snowflake data loads automatically using Amazon SQS (Simple Queue Service) notifications for an S3 bucket.
* Extensively wrote Postgres triggers for automating the ETL process in Postgres Database.
* Extracted the data from legacy systems into staging area using ETL jobs & SQL queries
* Perform Quality assurance testing and automated the error record detection on Postgres.
* Worked closely with the developers in API development using node.js in express frame work and wrote a service for data obfuscation.
* Performed unit and Integration testing on different API components using mocha and chai.
* Utilized google analytics to track the visitor flow and interaction throughout the company website.
* Linked the Jupiter note books on my local to the google analytics platform to perform analysis on the customer interaction data.
* Extensively researched and implemented various regression, classification and clustering Machine Learning algorithms in Jupyter notebooks.
* Created customer analytics metrics dashboard on google cloud platform using Big Query
* Strategic expertise in design of experiments, data collection, data analysis and visualization using various tools and technologies.

Environment: Apache Hadoop, EC2, ELB, VPC, S3, CloudFront, IAM, RDS, Route 53, AWS CloudWatch, SNS, AWS Lambda, AWS Glue, AWS SageMaker, AWS Personalize, RedSHift, Python, Maven, GIT, MySQL, PostgreSQL, Oozie, Sqoop, Flume, JDK 1.8, Agile and Scrum Development Process, google Analytics, Big Query, Dialog Flow.

**Role: Hadoop Developer Aug 2015 – Sep 2016**

**Company: CADSYS LIMITED, HYDERABAD, INDIA**

Responsibilities

* Analyzed Hadoop cluster using different big data analytic tools, including Pig, Hive, HBase, and MapReduce.
* Extracted data of everyday transactions of customers from DB2, exported it to Hive, and set up online analytical processing.
* Installed and configured Hadoop, MapReduce, and HDFS clusters.
* Created Hive tables, loaded the data, and performed data manipulations using Hive queries in MapReduce Execution Mode.
* Developed MapReduce programs to cleanse the data in HDFS obtained from heterogeneous data sources to make it suitable for ingestion into Hive schema for analysis.
* Loaded the structured data that resulted from MapReduce jobs into Hive tables.
* Analyzed user request patterns and implemented various performance optimization measures, including but not limited to implementing partitions and buckets in HiveQL.
* Identified issues in behavioral patterns and analyzed the logs using Hive queries.
* Analyzed and transformed stored data by writing MapReduce or Pig jobs based on business requirements.
* Used Flume to collect, aggregate, and store the web log data from different sources like web servers, mobile devices, and network devices, importing it into HDFS.
* Developed workflows using Oozie to automate data loading into HDFS and preprocessing with Pig scripts.
* Integrated MapReduce with HBase to import bulk data using MR programs.
* Utilized Maven extensively for building jar files of MapReduce programs and deployed them to the Cluster.
* Worked on developing Pig Scripts for data capture change and delta record processing between newly arrived data and already existing data in HDFS.
* Developed data pipelines using Sqoop, Pig, and Java MapReduce to ingest behavioral data into HDFS for analysis.
* Installed the Oozie workflow engine to run multiple Hive and Pig jobs, which run independently with time and data availability.
* Used Pig as an ETL tool to do Transformations, even joins, and some pre-aggregations before storing the data onto HDFS.
* Used SQL queries, Stored Procedures, User Defined Functions (UDF), and Database Triggers, using tools like SQL Profiler and Database Tuning Advisor (DTA).
* Installed a cluster, commissioned & decommissioned data nodes, performed name node recovery, capacity planning, and slots configuration, adhering to business requirements.