GOPI

E-Mail ID: gopichand.bigdtech@gmail.com **Contact No.: 571-250-7965**

Professional Summary:

- Having **10+** years of professional Software Development experience with specialization in Java/J2EE technologies and Big Data Analytics.
- Having **10+** years of work experience in ingestion, storage, querying, processing and analysis of Big Data with hands on experience in **Hadoop** Ecosystem including **MapReduce**, **HDFS**, **Hive**, **Pig**, **Spark**, **HBase**, **ZooKeeper**, **Sqoop**, **Flume** and **Oozie**.
- **Certified GCP and AWS Engineer**.
- Strong experience working with different Hadoop distributions like **Cloudera**, **Hortonworks**, **and AWS EMR,GCP**.
- In depth understanding/knowledge of **Hadoop** Architecture and various components such as **HDFS**, **MR**, **Hadoop GEN2** Federation, High Availability and **YARN** architecture and good understanding of workload management, scalability and distributed platform architectures.
- Experience in installing, configuring, supporting and managing **Hadoop** Clusters both on prem (**CDH**) and in cloud (**AWS EMR**), **GCP.**
- Experience in working with various big data analytics services in AWS like **EMR**, **S3**, **Redshift Spectrum**, **Athena**, **Glue** etc.,
- Strong experience building production ready **spark** applications, **troubleshooting** various failures in spark and **fine tuning** spark applications.
- Worked on **Google cloud platform** (GCP) services like **compute engine**, **cloud load balancing** ,**cloud storage**,**cloud SQL**,stack driver monitoring and cloud deployment manager
- Extensive experience working with **Spark RDD API's, Spark SQL** and **Spark ML**.
- Strong experience and knowledge of real time data analytics using **Kafka**, **Flume** and **Spark-Streaming**.
- Experience in extending **HIVE** and **PIG** core functionality by using custom **UDF**'s.
- Debugging **MapReduce** jobs using Counters and **MRUNIT** testing.
- Good understanding of Spark ML algorithms such as **Classification**, **Clustering**, and **Regression**.
- Experienced in moving data from different sources using **Kafka** producers, consumers and pre-process data using **Storm** topologies.
- Experienced in migrating **datawarehousing workloads** into **Hadoop based datalakes** using MR,Hive, Pig and Sqoop.
- Good knowledge on streaming data from different data sources like Log files, **JMS**, applications sources into **HDFS** using **Flume** sources.
- Experience in importing and exporting data using **Sqoop** from **HDFS** to Relational Database Systems and vice-versa.
- Extensive experience working with relational databases like Teradata, Oracle, Netezza, **SQL**Server and **MySQL** database.
- Excellent understanding and knowledge of NOSQL databases like MongoDB, **HBase**, and **Cassandra**.
- Worked on **Docker** based containerized applications.
- Knowledge of data warehousing and ETL tools like **Talend**, Power BI
- Experienced in working with monitoring tools to check status of cluster using **Cloudera** manager, **Ambari.**
- Experience with Testing MapReduce programs using **MR Unit**, **Junit**.
- Experience in design and development of Web forms using Spring MVC, Java Script, JSON and JQ plotter.
- Experience working on Version control tools like **SVN** and **Git** revision control systems such as **GitHub** and **JIRA/MINGLE** to track issues and crucible for code reviews.
- Worked on various Tools and IDEs like **Eclipse**, IBM Rational, **Visio**, Apache Ant-Build Tool, MS-Office, **PLSQL** Developer.
- Expertise in Azure data services, including **Azure** Data Factory, Azure Databricks, Azure SQL Data Warehouse, and Azure Cosmos DB.
- Implemented data pipelines using **Azure Data Factory**, ensuring efficient data ingestion, transformation, and loading processes.

- Designed and implemented data migration strategies to seamlessly transition data from legacy systems to new database structures, ensuring minimal disruption to operations.
- Stayed up-to-date with emerging technologies, trends, and best practices in **database design**, **access patterns**, and **API** development, actively seeking opportunities to apply new techniques to enhance system performance and scalability.
- Experience in different application servers like **JBoss/Tomcat**, Web Logic, IBM WebSphere.
- Experience in working with Onsite-Offshore model.

Technical Skills:

Big Data Ecosystem	**Hadoop, MapReduce, Pig, Hive, YARN, Kafka, Flume, Sqoop, Impala, Oozie, ZooKeeper, Spark.**
Hadoop Distributions	**Cloudera, AWS EMR, Hortonworks**
Languages	**Java, Scala, Python, SQL, Shell Scripting**
No SQL Databases	**Cassandra and HBase**
IDE	**Eclipse, IntelliJ**
Database	**Oracle 10g, MySQL, MSSQL**
Operating systems	**UNIX, LINUX, Mac OS and Windows Variants**
AWS Services	**AWS EMR, EC2, S3,Redshift,Athena,Glue**
GCP Servcies	**Cloud Storage,Big Query, Compute Engine, Cloud Functions**

Education Details:

- Bachelor of Engineering
- Masters in Computer Science

Certifications:

AWS Certified (ACTIVE)
Credential ID AWS02024790

GCP Certified (ACTIVE)
Credential ID vGr0T0
https://www.credential.net/d60b409e-fc42-4b4d-937a-0904c0836589?key=ffab6f33199d23b4c6ebc348340fdb399d46a1e0f63cad17dffe0c3fe9ab1e20

Publications:

1. Cloud-Native Operationalization of LLMs for Financial Compliance: A Managed AI Platform Approach
2. Metadata-Driven Self-Optimizing ETL Framework with Machine Learning-Based Workload Prediction
3. Metadata-Driven Self-Optimizing ETL Framework with Machine Learning-Based Workload Prediction
4. Lifecycle Management of Large Language Models in Financial Systems: Monitoring, Drift Detection, and Governance
5. Mitigating Bias and Data Poisoning in Large Language Model–Based Fraud Detection Pipelines

Professional Experience:

Annslo Tech Inc (WellsFargo Bank) **Dec.19 – Present.**
Sr. Software Engineer

Responsibilities:

- Involved in creating **data ingestion** pipelines for collecting health care and providers data from various external sources like FTP Servers and S3 buckets.

- Worked on **Google cloud platform** (GCP) services like **compute engine**, **cloud load balancing** ,**cloud storage**,**cloud SQL**,stack driver monitoring and cloud deployment manager.
- Involved in migrating existing traditional ETL jobs to **Spark** and **Hive** Jobs on new cloud data lake.
- Wrote complex **spark** applications for performing various denormalization of the datasets and creating a unified data analytics layer for downstream teams.
- Setup **GCP firewall** rules to allow or deny traffic to and from the VM's instances based on specified configuration.
- Good understanding of Data Ingestion , **Airflow** operators for Data **orchestration** and other related python libraries.
- Experience creating and managing **Airflow** DAGs
- Primarily responsible for **fine-tuning** long running spark applications, **writing custom spark udfs**, **troubleshooting** failures etc.
- Configured and optimized **Airflow DAGs** for efficient task execution, utilizing techniques such as task parallelization and dynamic task generation.
- Design star schema in **Big query**. Wrote a scala program for spark transformation in Dataproc.
- Involved in building a real time pipeline using **Kafka** and **Spark streaming** for delivering event messages to downstream application teams from an external rest-based application.
- Involved in creating **Hive scripts** for performing ad hoc data analysis required by the business teams.
- Written Spark applications to performing data cleansing, transformations, aggregations, and other useful datasets as per downstream team requirements using **Scala** and **PySpark**
- **Designed** and implemented **database tables** and schemas to optimize data storage and retrieval, ensuring efficient and scalable data management.
- Responsible for developing data pipelines involving ingesting raw JSON files, transactional and user profile information from on prem data warehouses and processing them using spark and finally loading the processed data to Big Query.
- Developed **access patterns** for the database, including defining indexes, **query optimization**, and data partitioning **strategies**, resulting in improved performance and reduced response times.
- Architected and implemented a solution to migrate the data platform from Hadoop to Google cloud Platform.
- Designed and developed **APIs** to expose data and functionality to internal and external systems, following industry best practices and standards.
- Collaborated with cross-functional teams, including software developers, product managers, and stakeholders, to gather requirements and design effective database structures and **access patterns** that aligned with business needs.
- Good experience with continuous integration of application using Jenkins.
- Used Cloud functions, Cloud pub-sub and Big Query to build streamline dashboard to monitor services.
- Created jobs using Cloud composer (**Airflow** DAG) to migrate data from Data Lake (cloud storage) to transform it using Data Proc and ingest in Big Query for further analysis.
- Automated resulting scripts and workflow using **Apache Airflow** and shell scripting to ensure daily execution in production.
- Designed and developed complex ETL workflows using **Apache Airflow**, resulting in improved data processing efficiency and reduced errors.
- Experienced in handling large datasets using **Spark** in-memory capabilities, using broadcasts variables in Spark, effective & efficient joins, transformations, and other capabilities.
- Deployed configuration management and provisioning Infrastructure and resources to GCP Cloud using Terraform and automated complete deployment environment in GCP.
- Utilized Google Cloud Storage as data lake and ensured all the processed data is written to **GCS** directly from spark and hive jobs.
- Setup **GCP firewall** rules to allow or deny traffic to and from the VM's instances based on specified configuration.
- Conducted thorough analysis of existing databases, identifying areas for improvement and proposing solutions to enhance data integrity, security, and performance.
- Implemented **data modeling techniques** to ensure data consistency, **normalization**, and referential integrity within the database.
- Worked closely with frontend and backend developers to define **API** specifications and data contracts, facilitating seamless integration between different components of the system.
- Primarily responsible for **fine-tuning** long running spark applications, **writing custom spark udfs**, **troubleshooting** failures etc.
- Hands on experience in setting up workflow using **Airflow** for managing and scheduling Hadoop Jobs.

- Design star schema in **Big query**. Wrote a scala program for spark transformation in Dataproc.
- Involved in building a real time pipeline using **Kafka** and **Spark streaming** for delivering event messages to downstream application teams from an external rest-based application.
- Good understanding of Data Ingestion , **Airflow** operators for Data **orchestration** and other related python libraries.
- Experience creating and managing **Airflow** DAGs
- Primarily responsible for **fine-tuning** long running spark applications, **writing custom spark udfs**, **troubleshooting** failures etc.
- Configured and optimized **Airflow DAGs** for efficient task execution, utilizing techniques such as task parallelization and dynamic task generation
- Solved performance issues in **Hive** and **Pig** scripts with understanding of Joins, Group and Aggregation and how does it translate to MR jobs.
- Design star schema in **Big query**. Wrote a scala program for spark transformation in Dataproc.
- Involved in building a real time pipeline using **Kafka** and **Spark streaming** for delivering event messages to downstream application teams from an external rest-based application.
- Involved in creating **Hive scripts** for performing ad hoc data analysis required by the business teams.
- Worked extensively on migrating on prem workloads to AWS Cloud.
- Worked on utilizing AWS cloud services like **S3**, **EMR**, **Redshift**, **Athena** and **Glue** Metastore.
- Used broadcast variables in spark, effective & efficient Joins, caching and other capabilities for data processing.
- Involved in continuous Integration of application using **Jenkins**.

Environment: AWS EMR, Spark, Hive, HDFS, Sqoop, Kafka, Oozie, HBase, Scala, MapReduce.

Client: Chick Fil A, Atlanta, GA **Oct'13 – Sep'19.**
Spark/Hadoop Developer

Responsibilities:

- Ingested click stream data from external servers such as FTP server and **S3** buckets on daily basis using custom Input Adapters.
- Created **Sqoop** scripts to import/export user profile data from RDBMS to **S3** data lake.
- Developed various **spark** applications using **Scala** to perform various enrichments of user behavioral data (click stream data) merged with user profile data.
- Involved in data **cleansing**, event **enrichment**, data **aggregation**, **de-normalization** and data **preparation** needed for down stream machine learning and reporting teams.
- Utilized **Spark** Scala **API** to implement batch processing of jobs.
- Troubleshooting **Spark** applications for improved error tolerance.
- **Fine-tuning spark** applications/jobs to improve the efficiency and overall processing time for the pipelines.
- Experience in **GCP Dataproc**, GCS, **Cloud Functions**, **Bigquery**.
- Build data pipelines in airflow in **GCP** for ETL related jobs using different airflow operators.
- Used cloud shell SDK in GCP to configure the services **Data proc ,storage ,BigQuery**.
- Co-ordinated with team and Developed framework to generate Daily adhoc reports and extracts from enterprise data from **BigQuery**.
- Created Kafka producer API to send live-stream data into various **Kafka** topics.
- Developed **Spark-Streaming** applications to consume the data from Kafka topics and to insert the processed streams to **HBase**.
- Utilized **Spark** in Memory capabilities, to handle large datasets.
- Used Broadcast variables in Spark, effective & efficient Joins, transformations and other capabilities for data processing.
- Experienced in working with **EMR** cluster and **S3** in AWS cloud.
- Worked with stakeholders to define requirements and scope for new **Airflow** projects, managing project timelines and deliverables to ensure successful implementation and adoption.
- Creating **Hive** tables, loading and analyzing data using hive scripts. Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
- Involved in continuous Integration of application using Jenkins.
- Interacted with the infrastructure, network, database, application and BA teams to ensure data quality and availability

- Working experience in **Agile** methodology

Environment: AWS EMR, Spark, Hive, HDFS, Sqoop, Kafka, Oozie, Hbase, Scala, MapReduce