**Bhanoji**

chinna@artemisinfotech.com | 281-954-2591

**Professional Summary\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* Boasting 9+ years of experience and a strong background with crafting data-rich applications and resolving data-driven business issues across different sectors encompasses hands-on experience in **Big Data** technologies as well.
* Proficient in utilizing **Big Data** integration technologies, including **GCP, Amazon Web Services (AWS), Microsoft Azure, Cassandra, HIVE, NoSQL** databases (such as **HBase** and **MongoDB**), and **SQL** databases (including **Oracle, SQL Server**, **Postgres SQL, MySQL Server, and Snowflake**).
* Extensive practical experience in utilizing **Amazon EC2, Amazon S3, Amazon RDS, Amazon Elastic Load Balancing, Amazon SQS,** AWS Identity and Access Management, **Redshift,** and various other AWS services.
* Accomplished development and deployment of multiple **Lambda functions** within AWS, utilized built-in AWS Lambda Libraries.
* Proficient in data analysis with Big Data Ecosystem, encompassing **HDFS, Hive, HBase, Zookeeper, PIG, Sqoop, and Flume**.
* Solid grasp of **Apache Airflow**, Proficient in orchestrating workflows using **Airflow, Data Pipelines, SSIS**.
* Demonstrated expertise in migrating SQL databases to **Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse**. Skilled in managing database access and permissions.
* Hands-on experience in batch and streaming data using **Spark Core APIs**, Spark Data Frames, Spark SQL, DStreams and spark structured streaming APIs.
* Worked on data lakes and processed huge volumes of structured, semi-structured and unstructured data.
* Skilled in developing **Spark applications** within **Databricks**, utilizing **Spark-SQL** for data extraction, transformation, and aggregation. Proficient in handling various file formats and analyzing data to reveal valuable insights into customer usage patterns.
* Comprehensive grasp of **Spark Architecture**, encompassing **Spark Core, Spark SQL, Data Frames, Spark Streaming,** Driver Node, Worker Node, Stages, Executors, and Tasks.
* Experienced on Cloud based stacks and Processed data using AWS EMR, Azure Databricks, Azure Synapse Analytics.
* Experienced in Data warehousing for Data Extraction, Data Transformation, Data Loading (ETL), Data Achieving, Purging, Masking, Data lineage and Data Analytics.
* Experience in performance tuning of the downstream jobs and calculated benchmarks over in-memory components.

**Technical Skills\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| Big Data Ecosystem | HDFS, MapReduce, Pyspark, Hive, Airflow, Sqoop, HBase. |
| Hadoop Distributions | Microsoft Azure - Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse. Amazon AWS - EMR, EC2, EBS, RDS, S3, Athena, Glue, Elasticsearch, Lambda, SQS, DynamoDB, Redshift, ECS.  Apache Hadoop 2.x/1.x |
| Scripting Languages | Python, JavaScript, R, PowerShell Scripting, HiveQL. |
| Cloud Environment | Amazon Web Services (AWS), Microsoft Azure. |
| No SQL Database | DynamoDB, HBase |
| Database | MySQL, Oracle, Teradata, MS SQL SERVER, |
| ETL/BI | Snowflake, SSIS, Power BI |
| Version control | Git, Bitbucket |
| Others | Jupyter Notebook, Jenkins, Jira, Airflow |

**work experience**

**Client:** Incomm Jan 2023 – Till Date

**Location:** Alpharetta, GA

**Role: Sr. Data Engineer**

**Responsibilities:**

* Collaborated closely with business stakeholders to gather requirements for data warehousing and reporting.
* Performed data extraction, transformation, and loading tasks from diverse sources into **Azure Data Storage Services**, utilizing **Azure Data Factory** and **t-SQL** for data lake analytics.
* Executed data transformations essential for ML Operations, includes tasks such as adding calculated columns, managing relationships, creating measures, merging, and appending queries, value replacement, column splitting, grouping by, and handling Date & Time data.
* Managed data ingestion into various **Azure services**, including **Azure Data Lake, Azure Storage, Azure SQL, and Azure Data Warehouse**, while efficiently processing data within **Azure Databricks**.
* Established scheduled batches and sessions for data movement, both at specified intervals and on-demand, through Server Manager.
* Integrated multiple data connections, established joins across data sources as part of data preparation efforts.
* Developed Python scripts within Databricks, performing file validations and automated process through ADF.
* Worked with **Pyspark**, improving the performance, optimized the existing applications running on **Data Bricks**.
* Worked on **Databricks** and **snowflake integration** and processed data. Developed and Scheduled **Databricks** notebooks using Autoloader – Delta live tables.
* Worked on Data lakes and handled huge volumes of data using Spark SQL, data frames and measured benchmarks with Hive-Tez.
* Worked on **DBT - Databricks** for ELT loads and Experience on dela loads.
* Generated tables within **Azure SQL Data Warehouse** to support data reporting and visualization in alignment with specific business requirements.
* Produced **visualization reports, dashboards, and KPI** scorecards utilizing **Power BI** Desktop.
* Designed, developed, and deployed **ETL** solutions using **SQL Server Integration Services (SSIS).**
* Experience with **Snowflake** Virtual Warehouses, worked on Dimensional and Relational Data Modelling using Star and **Snowflake** Schemas, OLTP/OLAP system, Conceptual, Logical and Physical data modelling.
* As a part of Data Migration, wrote many SQL Scripts for Mismatch of data and worked on loading the history data from Teradata SQL to **snowflake**.
* Exported Data into **Snowflake** by creating Staging Tables to load Data of different files from Amazon S3. Scheduled daily running jobs to build and monitor regularly in **Data-Lineage**, configured email alerts whenever pipeline fails.
* Designed and Developed ETL Processes in **Data Bricks** to migrate Campaign data from external sources like Azure Data Lake, gen2 in Parquet/Text Files.
* Experienced in loading and transforming large sets of structured, semi-structured and unstructured data. Developed **Spark** jobs and Hive Jobs to summarize and transform data.
* For Query optimization and fast query retrieval performed Normalization and De-Normalization of existing tables, with the effective use of Joins & indexes.
* Creating alerts on data integration events (success/failure) and monitoring them. Collaborating with product managers, scrum masters, and engineers to develop Agile practices and documentation initiatives to bring experience for retrospectives and meetings.

**Client:** Horizon Media Oct 2021 - Dec 2022

**Location:** New York, NY

**Role: Data Engineer**

**Responsibilities:**

* Designed and implemented a data pipeline for automating ingestion, processing, and delivery of both batch and **streaming data, employing Spark,** **AWS EMR Clusters**, **Lambda, and Databricks**.
* Developed **Airflow workflows and Python scripts** to handle batch data processing, **ETL** tasks, and data warehouse ingestion, utilizing **AWS Lambda Python functions, Elastic Kubernetes Service (EKS), S3.**
* Managed data ingestion into a **data lake (S3)** and leveraged **AWS Glue** to expose this data to **Redshift**.
* Configured EMR clusters to facilitate data ingestion and utilized dbt (data build tool) for data transformation within Redshift.
* Developed, tested, and optimized **NoSQL** code for transformations using the data build tool.
* Constructed data pipelines with **Airflow** to schedule **PySpark** jobs for incremental loading, and integrated Flume for processing weblog server data.
* Automated data ingestion into data lake through scheduled Airflow jobs using Apache Airflow within a cluster environment.
* Scripted Python code and employed **Airflow DAGs** to automate the extraction of weblogs.
* Extensive experience in automation and deployment with like **AWS** **IAM**, **AWS EMR** and **Redshift**
* Assigned Lambda function with **IAM Role** with necessary permission to push from the source bucket to the destination bucket.
* Worked with **AWS** **IAM** in managing technologies such as vulnerability assessment tools, identity and access management, web content filtering, VPN / two factor authentication planning and solutions.
* Designed scalable applications for real-time data ingestion into various databases using **AWS Kinesis**, performed necessary transformations, aggregation to common learner data model, storing data in HBase.
* Orchestrated multiple **ETL** jobs using **AWS Step Functions and Lambda**, while leveraging **AWS Glue** to load and prepare data for analytics.
* Utilized **AWS Lambda** to manage serverless computing, trigger code execution via **S3** and **SNS** events, and developed data transition programs from **DynamoDB** to **AWS Redshift** through Python functions.
* Effectively implemented **AWS cloud computing**, integrating **RDS, Python, DynamoDB, S3, and Redshift**.
* Developed Spark applications using **Spark-SQL** in **Databricks** for data extraction, transformation, and aggregation across various file formats, enabling insights into customer usage patterns.
* Worked with diverse file formats, including delimited **text files, clickstream logs, Apache log files, Avro files, JSON files, and XML file**s.
* Demonstrated expertise in utilizing different columnar file formats such as **RC, ORC, and Parquet**.
* Created and implemented **NoSQL** queries, views stored procedures and triggers.
* Collected near-real-time data from **AWS S3 buckets** using **Spark Streaming** and executed on-the-fly transformations to build a common learner **data model**.
* Setup Spark EMR to process huge data which is stored in **AmazonS3** and used **Amazon CLI** for data transfers to and from Amazon S3 buckets.
* Provided support for production data pipelines, including performance optimization, and troubleshooting of **SQL, Spark, and Python scripts**.

**Client: Intuit** Jan 2020 - Oct 2021

**Location: Plano, Texas**

**Role: Spark Developer**

**Responsibilities:**

* Played a key role in migrating existing on-premises Hive code to **GCP (Google Cloud Platform) BigQuery**.
* Contributed to the migration of an **Oracle SQL ETL** process to operate on Google Cloud Platform, utilizing Cloud Data Processing and **BigQuery**. Leveraged **Cloud Pub/Sub** for triggering Apache Airflow jobs.
* Designed and developed applications for data processing using Spark.
* Designed and implemented **Hive scripts, and PIG scripts** as part of a Data Warehouse migration project.
* Designed system for collecting data from multiple portals using **Kafka**, processing it efficiently with Spark.
* Collected data using **Spark Streaming** from AWS S3 bucket in near-real-time and performs necessary transformations and Aggregations to build the data model and persists the data in HDFS.
* Automated data movement and purging processes using shell scripting.
* Expertise in implementing **Spark Scala** application using higher order functions for both batch and interactive analysis requirement. Authored Hive jobs to parse and structure logs into a tabular format to enhance log data querying.
* Managed data import and export between **Impala, HDFS, and Hive** using **Sqoop**. Oversaw management of data from various sources.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark data frames**, Scala, and Python.
* Created **Hive** tables for data transformation and analysis within **HDFS**.
* Implemented **Spark** using **Scala** and utilizing Data frames and **Spark SQL API** for faster processing of data.
* Participated in the creation of **Hive tables, data loading**, and the development of internal **Hive queries**.
* Executed **Hadoop Jobs** for processing large volumes of text data records.
* Expertise in implementing Spark Scala application using higher order functions for both batch and interactive analysis requirement.
* Established **JDBC** connections and employed **JDBC** statements to invoke stored procedures.
* Developed Pig Latin scripts for data extraction from web server outputs, preparing data loading into HDFS.
* Designed Pig **UDFs** to preprocess data for analysis.
* Implemented multiple **Java**-based Map Reduce Jobs for data cleansing and preprocessing.
* Transferred **RDBMS** data into flat files from various channels, loaded into **HDFS** for further processing.
* Created job workflows in Oozie to automate tasks related to data loading into **HDFS**.

**Client: Novartis** July 2017 - Dec 2019

**Location: Parsippany, New Jersey**

**Role : Hadoop Developer**

**Responsibilities:**

* Authored script files to facilitate data processing and loading into **HDFS**, conducted data analysis leveraging **Pig, Hive**, and delivered summarized insights from Hadoop to downstream systems.
* Employed Pig as an ETL tool to perform data transformations, event joins, and preliminary aggregations before persisting data on **HDFS**.
* Established a data pipeline utilizing **Flume, Sqoop, and Pig** to extract and store weblogs data in **HDFS**.
* Utilized **Sqoop** for seamless data import and export between **HDFS** and relational databases (**RDBMS**).
* Orchestrated export of analyzed data to **MySQL** relational database using Sqoop for visualization and report generation.
* Constructed **HBase** tables to handle substantial volumes of structured data.
* Oversaw and scrutinized **Hadoop** log files to ensure system health and performance.
* Contributed insights for estimation in new proposal preparation.
* Extensively interacted with **HIVE Data Definition Language** **(DDL),** employed Hive Query Language **(HQL)** for data operations.
* Designed User-Defined Functions (UDF), User-Defined Aggregate Functions (UDAF), and User-Defined Table-Generating Functions (UDTF) to enhance Hive queries.
* Implemented **Sqoop** for seamless large dataset transfers between **Hadoop** and RDBMS systems.
* Leveraged Sqoop extensively to import data from diverse systems/sources (e.g., MySQL) into **HDFS**.
* Developed Hive **User-Defined Functions (UDFs)** to address missing functionality for analytics.
* Managed cluster coordination services; Worked with various file formats, Text files, Sequence Files, and Avro.
* Responsible for building scalable distributed data solutions using **Hadoop**.
* Collaborated development and technical documentation for launching **Hadoop** clusters, executing Hive queries and Pig scripts.
* Provided support in cluster maintenance, monitoring, node management (additions and removals), and troubleshooting.
* Successfully installed and configured **Hadoop** and **HDFS**, created multiple jobs for data cleansing and pre-processing.
* Assisted in creating and maintaining technical documentation to launch **HADOOP** Clusters and even for executing Hive queries and Pig Scripts.
* Assisted in Cluster maintenance, cluster monitoring, adding, and removing cluster nodes and Trouble shooting.

**Client: Concentrix technology** June 2014 - June 2017

**Location: India**

**Role: DATA ANALYST**

* Leveraged **Python**, **SQL** to analyze substantial **3TB** data set for comprehensive revenue audits of 165000 clients in Comerica bank.
* Achieved cost savings of **$24k/yr** for team by automating **22** integrated **PowerBI** dashboards, derivation of target metrics and KPIs
* Enhanced client efficiency by **25%** through data-driven decision making of financial data reports using **PowerBI** dashboards.
* Generated a **12%** revenue boost through innovative statistical models, ascertain the influence of various factors on loan defaulters.
* Collaborated with senior analysts, refined regression analysis-derived insights and achieved a **15%** reduction in marketing expenses
* Attained **90%** reduction in data inconsistencies, saved over **15hr/week** to streamline data cleaning by automation of **Python** scripts
* Developed data pipelines for ingesting, processing, and transforming large datasets, resulting in a **20%** increase of data availability.
* Integrated **SharePoint** with tools like **PowerBI**, **AWS Redshift**, to facilitate smooth data flow and enhance reporting capabilities.
* Optimized ETL processes, enhanced data workflow efficiency and achieved **30%** reduction in data processing times.
* Created detailed ad hoc reports highlighting significant outcomes for senior leadership, impacting banking strategy and operations.
* Managed **Tableau** to **QuickSight** migration, **$15k** annual savings, reduced reporting time by **30%** via Python-integrated dashboard
* Created a Python-based dynamic math model, implemented metadata in Data Lake using Athena, SQL, and MS-Excel to optimize analytical processes and RCA.