**Anil Mandava**

mailanilmandava@gmail.com | +1 (330)-969-9139

**Summary:**

* Around **9+** years of experience in the IT industry with expertise in design and development using **Java, Hadoop, Spark.**
* Extensive experience in Big Data ecosystem and its various components **MapReduce, Spark Sql, SQL, HDFS, Hive, HBase, PIG, Sqoop, Zookeeper, Oozie, Airflow**
* Strong knowledge on Hadoop architectural components like Hadoop Distributed File System, Name Node, Data Node, Task Tracker, Job Tracker, MRV2, and Map Reduce programming.
* Implemented **Spark SQL and Data frame API** to connect to Hive to read the data and distributed processing to make highly scalable.
* Experience in writing MapReduce ingestion and transformation jobs using **Sqoop, Pig, Hive** for data processing.
* Exposure on NoSQL databases and hands on work experience in writing applications on **NoSQL databases like HBase and Casandra**.
* Knowledge in real time streaming applications and batch style large scale distributed computing applications using tools like **Spark Streaming**.
* Experience in deploying Hadoop cluster on **Amazon Web Services (AWS) EMR clusters.**
* Experienced on **GCP** services GCS, Virtual Engines, Cloud SQL, BigQuery.
* Experience in working with **distributed scheduling using Oozie and Airflow**.
* Extensive knowledge of **Snowflake SQL**, with experience in creating complex database joins, optimizing queries, and designing efficient databases.
* Excellent knowledge of job workflow scheduling and locking tools/services like **Oozie and Zookeeper.**
* Experienced in utilizing cloud-based data warehousing solutions like **Snowflake**, and proficient in developing and maintaining data pipelines from various sources to the data warehouse.
* Good understanding of Partitions, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Hands-on experience in migrating on-premises ETL processes to **Google Cloud Platform (GCP)** using cloud-native tools like **BigQuery, Cloud DataProc, Google Cloud Storage,** and **Composer**.
* Expertise in building real-time **distributed stream processing pipelines** using technologies such as **Spark, Kafka,** and **NiFi**.
* Expertise in writing automation using shell scripting and Java.
* Capable of processing large sets of **structured, semi-structured and unstructured data** and supporting systems application architecture.
* Closely working with Data analysts and business users by providing with normalized data to help with different use-cases.
* Worked with various formats of files like delimited text files, **JSON** files, **XML** Files. Mastered in using different columnar file formats like **RC**, **ORC** and **Parquet** formats and has a good understanding of various compression techniques used in Hadoop processing like **G-zip**, **Snappy**, **LZO** etc.
* Experience in the data analysis and debugging skills.
* Experience in implementing dynamic web applications using **Apache Tomcat 6.x**
* Strong knowledge in working with MVC architecture using Struts, Spring Framework.
* Used Git for version control and Jira for project management, tracking issues and bugs.
* Knowledge on Application Development and maintenance life cycle process.

**Technical Skills:**

**Big data technologies:** Hadoop, HDFS, Spark, Hive, Sqoop, Map-Reduce, Oozie, pig, HBase, Yarn, Flume

**Data base and Data warehouse:** Oracle, MS-SQL Server, MySQL, Teradata, Snowflake.

**NOSQL:** HBase, Cassandra

**Programming languages:** Java, Python, HiveQL

**CI/CD Tools:** Jenkins, Maven, Git, Jira

**IDE &Tools:** Eclipse, Hue, IntelliJ, Oracle (PL/SQL), Putty, PyCharm

**Scripting Languages:** HTML, UNIX shell scripting, Java Script

**Hadoop Distributions:** Cloudera, Google Cloud Platform, AWS Services, Hortonworks

**Professional Experience:**

**Data Engineer – Walmart, Sunnyvale, CA Jan 2023 – Till date**

**Responsibilities:**

* Migrating on-prem ETLs from MS SQL server to GCS Cloud.
* Building Reusable Data ingestion and Data transformation frameworks using Python.
* Build, enhance, optimized Data Pipelines using Reusable frameworks to support data need for the analytics and Business team using Spark and Kafka.
* Worked on creating pipelines to ingest data from Hadoop to GCP BigQuery.
* Involved in building the ETL architecture and Source to Target mapping to load data into Data warehouse.
* Used ETL to develop jobs for extracting, cleaning, transforming, and loading data into data warehouse.
* Performed Spark jobs with the Spark core, SparkSQL libraries for Processing the data.
* Developed Spark Scripts by using Python, Shell commands as per the Requirement.
* Optimization of code logic in PySpark to reduce the time taken to run scripts and save developers’ time for other tasks
* Worked on reading and writing multiple data formats like JSON, ORC, Parquet on HDFS. using PySpark.
* Experience in Hive Partitioning, Bucketing and perform Joins on Hive tables.
* Developed Hive queries for data sampling and analysis to the analysts.
* Worked on formulating business metric queries on GCP BigQuery and validating across multiple datasets.
* Involved in writing the Shell scripts for exporting log files to Hadoop cluster through Automated process.
* Experience on preparing test data and executed the detailed Test plans. Completed required debugging.
* Created monitors, alarms, notifications, and logs for Data using Crontab.
* Worked with clients on Data Science project related to Household analysis and Customer churn rate in Walmart.
* Used Cronjobs End to End data processing pipelines and Scheduling the Workflows.
* Managed and implemented Daily continuous code deployments using Jenkins.
* Experienced in development actives in complete Agile model using JIRA and GIT.

**Data Engineer – Vanguard, Dallas, TX Jul 2022 – Jan 2023**

**Responsibilities:**

* Designed and developed a data platform which is consumed by downstream machine learning models for targeted prospect and client marketing.
* Ingested machine learning models into Adobe Audience Manager to identify and create audience profiles for personalized marketing campaigns.
* Developed Serverless ETL pipelines to automate the process of discovering data schema, categorizing data, publishing metadata, and automated by Airflow jobs to clean, enrich and move the data to DataLake for downstream analytics.
* Automated manual processes extensively and scheduled them using Python /Shell scripts.
* Converted SQL queries into Spark transformations using DataFrames and Python API.
* Tuned and optimized Spark job performance to prevent resource bottlenecks and reduced the job runtime from 1.30 hrs to 15 minutes while processing analytical workloads in AWS EMR.
* Designed validation rules to perform automated record level data quality & integrity checks.
* Improved data quality by working with data analysts and product owners to deploy data quality checks to detect and filter out 15%-20% low quality records.
* Designed and developed data pipelines to implement slowly changing dimensions (SCD1, SCD2) for client attributes like name, email, address etc...
* Reduced the time to process the daily loads by 50% by implementing Change Data Capture (CDC) and incremental loading.
* Created various SQL objects such as schema, tables, field definitions and mappings, CTE and views.
* Worked with different file formats like Text, Avro, ORC, and Parquet for HIVE or Presto querying and processing.
* Orchestrated and monitored batch transformation jobs performed on AWS EMR clusters.
* Created robust monitoring systems & notification alerts to handle runtime failures using AWS CloudWatch & SNS.
* Reduced the time to elevate the application code to production by creating CI/CD pipelines with Bamboo.

**Hadoop Developer – Infosys, Bangalore. April 2020 – Jan 2022**

**Responsibilities:**

* Involved in building real-time data ingesting pipelines using NiFi.
* Build CI/CD pipelines for deployment of packaged files.
* Worked on creating pipelines to ingest data from Hadoop to GCP BigQuery.
* Worked on formulating business metric queries on BigQuery and validating across multiple datasets.
* Involved in developing end to end Spark analytical applications for business insights.
* Involved in requirement gathering and designing the solution.
* Worked on getting structured data from Teradata to Hive, HDFS using TDCH.
* Worked on validating data after import.
* Created Oozie workflows to trigger Spark and MLP jobs.
* Developed Java application to trigger Oozie workflows and monitor the jobs’ status.
* Implemented Hibernate to log job details in MySQL.
* Developed Python scripts for coordination with different tools in the organization.
* Worked on optimizing Hive and Spark scripts.
* Worked on converting Hive queries to Spark using Python.
* Worked on Spark optimizations and memory management.
* Imported millions of structured data records from relational databases using Sqoop import to process.
* Created and worked Sqoop jobs with incremental load to populate Hive External tables.
* Developed mappings using data processor transformation to load data from word, pdf documents to HDFS.
* Solved performance issues in Hive scripts with understanding of joins, group and aggregations.
* Fetching the HQL results into CSV files and handover to reporting team.
* Collaborated with team engineers to produce high quality code using Agile software development.
* Worked on POC for streaming data using Kafka and Spark Streaming.

**Software Developer – Lince Soft Solutions Private Limited, India. Feb 2018 – Mar 2020**

**Responsibilities:**

* Worked on a live 30 node Hadoop cluster running Cloudera.
* Experienced with through hands-on experience in Hadoop, Java, SQL and Python.
* Importing and exporting data into HDFS from Oracle database and vice versa using Sqoop.
* Used Flume to collect, aggregate, and store the web log data from different web servers, network devices and pushed to HDFS.
* Load and transform large sets of semi-structured and unstructured data that includes Sequence files and XML files and worked on Avro and Parquet file formats using compression techniques like Snappy, Gzip and Zlib.
* Experienced working on processing unstructured data using Pig and Hive.
* Worked on Sequence files, RC files, Map side joins, bucketing, Partitioning for Hive performance enhancement and storage improvement.
* Consuming the data from web services using swagger UI for REST APIs.
* Experienced in writing the Map Reduce programs for analyzing of data as per the business requirements.
* Experienced in using Hive and Pig as an ETL tool for event joins, filters, transformations and pre- aggregations.
* Evaluated Oozie for workflow orchestration in the automation of MapReduce jobs, Pig and Hive jobs.
* Developed Junit test cases to validate the results of analysis of MapReduce.
* Experienced in managing and reviewing Hadoop log files.
* Improved and maintained legacy applications to extend its longevity.
* Involved in analyzing system failures, identifying root causes, and recommended course of actions.

**Java Developer –MicroSpark Software Solutions, India Feb 2014 – March 2018**

**Responsibilities:**

* Involved in design, development and analysis documents in sharing with Clients.
* Analysis and Design of the Object models using JAVA/J2EE Design Patterns in various tiers of the application.
* Developed Presentation Layer using HTML, CSS, and JSP and validated the data using JavaScript.
* Analyzing the Client Requirements and designing the specification document based on the requirements.
* Applied J2EE design patterns like business delegate, DAO and Singleton.
* Worked with Maven build tool to build the Project.
* Written SQL queries, PL/SQL and stored procedures as part of database interaction.
* Used dispatch action to group related actions into a single class.
* Testing and production support of core java based multithreading ETL tool for distributed loading XML data into Oracle10g database using JPA/Hibernate.
* Utilized frameworks such as Hibernate and Spring for persistence and Application Layers.
* Attached an SMTP server to the system, which handles Dynamic E-Mail Dispatches.
* Defined and Developed Action and Model Classes.
* Write unit tests using JUnit, Mockito to ensure the code produced is testable and reaches high  
  quality levels.
* Used Spring Framework and created the Dependency injection for the Action classes using ApplicationContext.xml.
* Configured and Deployed application on Tomcat Application Server.
* Used Log4j to implement logging facilities.
* Used Git for version control.