**CHANDRASEKHAR SIMMANA**

simmana.chandra@gmail.com • 240-308-4499 • Newark, DE • [www.linkedin.com/in/chandra-simmana/](https://www.linkedin.com/in/chandra-simmana-25167088/)

**EXECUTIVE PROFILE**

Results-driven IT professional with 11+ years of experience in designing, developing, and optimizing large-scale data applications. Expertise in Big Data, ETL, Hadoop ecosystem, Spark, and cloud-based solutions. Strong background in data modeling, migration, and performance optimization using various databases and cloud platforms. Proficient in automation, scripting, and system analysis within Agile and Waterfall methodologies. Adept at translating business needs into scalable data solutions, leveraging advanced analytics and reporting to drive strategic insights and decision-making.

**TECHNICAL SKILLS**

**Big Data & Cloud:** Hadoop, Cloudera (CDH/CDP), AWS (S3, Glue, Redshift, EC2, DynamoDB, Athena), Azure

**Programming & Scripting**: Python, Scala, Java, Shell Scripting

**Databases**: Oracle, SQL Server, Teradata, MongoDB, HBase, Hive

**ETL & Data Processing:** Spark, PySpark, Sqoop, Hive, Kafka, Airflow

**CI/CD & Version Control:** GitHub, Jenkins, Ansible, Control-M, Autosys

**PROFESSIONAL EXPERIENCE**

**Big Data Engineer (May 2024 – Present)**

SEI Investments Projects - Philadelphia, PA

* Designed and optimized scalable cloud-based data pipelines using Databricks on AWS, leveraging Spark for large-scale data processing.
* Developed ETL workflows integrating Databricks with AWS services (S3, Glue, Redshift, Lambda) for seamless data ingestion and transformation.
* Implemented Delta Lake to enhance data reliability, enabling ACID transactions, time travel, and efficient data storage for analytics.
* Utilized DBT (Data Build Tool) to transform and model data in Snowflake, creating reusable and modular data transformation pipelines for analytics and reporting.
* Optimized Spark jobs for performance tuning, cost efficiency, and resource management using Databricks clusters and AWS EMR.
* Automated workflows using Apache Airflow and Databricks Jobs, ensuring efficient orchestration and monitoring of end-to-end data pipelines.
* Built scalable and reliable ETL pipelines for processing large datasets, supporting both real-time and batch data processing systems.
* Designed data storage solutions using distributed databases, data warehouses, and data lakes to meet business analytics and reporting needs.
* Architected scalable and secure AWS infrastructures, collaborating with cross-functional teams to ensure seamless integration and minimal disruption.
* Integrated data from various structured and unstructured sources (APIs, databases, logs, streaming services) into centralized data platforms.
* Transferred and validated metadata from Hadoop ecosystems (e.g., Hive metastore) to AWS services, ensuring data consistency and accessibility.
* Executed data transfers from on-premises HDFS to AWS S3, maintaining data integrity and minimizing downtime.
* Designed fault-tolerant systems to manage cascading failures in Spark tasks, ensuring robust data processing.
* Automated deployment processes using CI/CD tools and implemented monitoring tools to track pipeline performance and data flow.

*Key Achievements:*

* Improved data processing efficiency by **30%** through optimization of **Spark jobs**, reducing runtime and resource costs.
* Automated **15+ workflows** using **Apache Airflow**, reducing manual intervention and improving pipeline reliability by **25%**.
* Implemented DBT to streamline data transformation processes in Snowflake, reducing development time for analytics pipelines by 20%.

**Platform Used:** Databricks

**Big Data Engineer (July 2022 – March 2024)**

Bank of America Corporate, Newark, DE

* Led on-prem data migration initiatives, transitioning legacy systems to a modern Hadoop-based platform to enhance operational efficiency and data integrity.
* Designed and implemented data ingestion pipelines using Sqoop to import/export data between HDFS and relational databases (e.g., Oracle, SQL Server, Teradata), ensuring seamless data integration.
* Developed Hive tables and wrote HiveQL queries for data analysis, leveraging partitioning and bucketing to optimize query performance.
* Utilized Impala for fast, interactive SQL queries on large datasets stored in HDFS, improving query response times by 30%.
* Implemented Kafka for real-time data streaming, enabling ingestion of high-velocity data from multiple sources into HDFS and Hive.
* Automated data workflows using Oozie, scheduling and orchestrating Hive, Pig, and Spark jobs for daily data processing and ETL tasks.
* Optimized Hive/Spark ETL jobs by tuning queries, reducing data skew, and leveraging partitioning, resulting in a 40% reduction in job execution time.
* Migrated Cloudera clusters from CDH 5.2/6.1 to CDP 7.1, ensuring compatibility and performance improvements across the Hadoop ecosystem.
* Monitored and managed Hadoop clusters using Cloudera Manager (CM), troubleshooting performance bottlenecks and ensuring high availability.
* Conducted cloud POCs for AWS and Azure, exploring potential migration strategies and evaluating cloud-based solutions for future scalability.
* Worked with structured and unstructured data, integrating data from RDBMS, flat files, and streaming sources into HDFS and Hive for centralized processing.
* Developed MapReduce programs for data cleaning and analysis, processing terabytes of data efficiently.

*Key Achievements:*

* Successfully migrated 10+ TB of data from on-premises HDFS to AWS S3, ensuring 99.9% data integrity and minimal downtime.
* Optimized 20+ Hive/Spark ETL jobs, reducing job execution time by 40% and saving $50K annually in computational costs.

**Platform Used:** Hadoop (HDFS, Hive, Impala, Sqoop, Oozie, Kafka), Spark, Cloudera (CDH/CDP), Oracle, SQL Server, Teradata

**Hadoop Developer (March 2021 – June 2022)**

AMFAM, Madison, WI

* Designed and implemented cloud-based data pipelines on AWS, leveraging S3, Redshift, and Lambda for data storage and processing.
* Developed Sqoop ingestion scripts to load data from RDBMS (e.g., Oracle, Teradata) into Hive and HDFS for daily data ingestion processes.
* Created and managed Hive tables, optimized queries using partitioning and bucketing, and wrote HiveQL for data transformation and analysis.
* Utilized Impala for fast, interactive SQL queries on large datasets, reducing query response times by 25%.
* Implemented Kafka for real-time data streaming, enabling ingestion of high-velocity data from multiple sources into HDFS and Hive.
* Automated Hadoop jobs using Apache Oozie, scheduling and orchestrating Hive, Pig, and Spark jobs for daily data processing.
* Developed MapReduce programs for data cleaning and analysis, processing terabytes of data efficiently.
* Worked with structured and unstructured data, integrating data from RDBMS, flat files, and streaming sources into HDFS and Hive for centralized processing.
* Monitored and managed Hadoop clusters using Cloudera Manager (CM), troubleshooting performance bottlenecks and ensuring high availability.

*Key Achievements:*

* Automated AWS snapshot creation using Python scripts, reducing manual effort by 50% and ensuring consistent backups.
* Designed and implemented a data warehouse on AWS Redshift, handling 5+ TB of data and improving query performance by 35%.

**Platform Used:** Hadoop (HDFS, Hive, Impala, Sqoop, Oozie, Kafka), Spark, Cloudera (CDH/CDP), AWS (S3, EC2, Redshift, Lambda), Oracle, Teradata

**Hadoop Developer (March 2019 – Feb 2021)**

Marico - Hyderabad, India

* Collaborated with architects to translate functional and technical requirements into detailed architecture and design.
* Validated transactional and profile data from RDBMS transformed and loaded into Data Lake using Hadoop technologies.
* Automated file copying processes in Hadoop for testing purposes and monitored server health using Splunk.
* Developed MapReduce and Hive scripts for data ingestion and validation, and converted Hive/SQL queries into Spark transformations using Spark RDDs.
* Processed schema-oriented and non-schema-oriented data using Scala and Spark, and collected data from multiple portals using Kafka and Spark.
* Designed and built a Reporting Application using Spark SQL to fetch and generate reports from HBase table data.
* Developed custom Hive UDFs and Pig scripts to transform large volumes of data and built scalable distributed data solutions using Hadoop.

*Key Achievements:*

* Developed and deployed 10+ custom Hive UDFs, reducing data transformation time by 20% and improving data processing efficiency.
* Automated daily Hadoop jobs using Apache Oozie, reducing manual effort by 30% and ensuring timely data availability for analytics.

**Software Developer (March 2016 – Feb 2019)**

Ericsson - Hyderabad, India

* Designed and implemented Hadoop-based applications to process large datasets, optimizing MapReduce programs, HiveQL queries, and Pig scripts.
* Built and managed ETL workflows using Apache Hive, Pig, and Spark, and developed data ingestion processes for structured and unstructured data.
* Leveraged Hadoop ecosystem tools (HDFS, MapReduce, Hive, HBase, Sqoop, Oozie, Flume) for data processing and analysis.
* Optimized Hadoop cluster configurations and integrated data from relational databases, flat files, APIs, and real-time streams.
* Implemented data validation, cleansing, and quality checks within Hadoop workflows and ensured data accuracy and reliability.
* Automated data workflows using Apache Oozie and Autosys, and wrote scripts in Python, Java, and Shell for automation tasks.
* Collaborated with data analysts, engineers, and business stakeholders to understand requirements and document development processes.

*Key Achievements:*

* Optimized 15+ MapReduce jobs, reducing job execution time by 25% and improving cluster resource utilization by 20%.
* Integrated 5+ data sources (relational databases, APIs, flat files) into Hadoop, enabling centralized data processing for 10+ business units.

**Software Developer (Sep 2012 – Feb 2016)**

Value IT Labs

* Participated in Object-Oriented Analysis and Design (OOAD) sessions for projects based on MVC architecture using the Spring Framework.
* Implemented Object-Relational Mapping (ORM) in the persistence layer using Hibernate and generated POJO classes for database table mapping.
* Applied Spring Transaction and Spring AOP methodologies for transaction management and wrote SQL queries and stored procedures for database communication.
* Conducted unit testing using the JUnit framework and built and deployed applications using Maven and SVN for version control.

*Key Achievements:*

* Reduced database query response time by 40% through optimization of Hibernate fetching strategies and SQL queries.
* Automated CI/CD pipelines using Maven and SVN, reducing deployment time by 30% and improving release cycle efficiency.

**EDUCATION**

**Bachelor’s in computer science engineering May-2012**

*JNTU University*