# 

# Chandra sai kiran kammari

**Data Engineer**

**Email: chandrasaikirankammari@gmail.com**

**Phone: +18139221811**

## Professional Summary:

* **10 years** of professional IT experience in designing and building **Data Management Lifecycle** including Data Ingestion, Data integration, Data consumption, Data delivery, and integration Reporting, Analytics, and System-System integration.
* Experience in evaluating the technology stack for building **Cloud-Based Analytics** solutions by conducting research and identifying appropriate strategies, tools, and methodologies for developing end-to-end analytics solutions and assisting in the development of a technology roadmap for **Data Ingestion, Data Lakes, Data Processing, and Data Visualization.**
* Good experience on Data Management, Analytics, performing analysis, design, development, testing, and implementation of large-scale On-premises and **Cloud SQL** databases and data warehouses using **Microsoft Business Intelligence Products and Azure**.
* Proficient in **Big Data development** with the Hadoop ecosystem including **Hadoop MapReduce, HDFS, Spark, Hive, HBase, Cassandra, Sqoop, Flume, Kafka, Zookeeper, Impala,** and **Oozie.**
* Extensive experience in **Spark Architecture** including Spark Core, Spark SQL, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors, and Tasks.
* Extensively used **Spark Data Frames API** over the **Cloudera platform** to perform analytics on Hive data and used Spark Data Frame Operations to perform required **data validation**.
* Strong experience in implementing the **Data Ingestion** techniques using the **Azure Data Factory** and **Databricks** in Azure cloud.
* Good experience on Azure services such as **Azure Data Lake, Azure Blob Storage, and Azure Data Bricks.**
* Expert in recreating existing application logic and functionality in the **Azure Data Lake, Data Factory, SQL Database, and SQL Data warehouse environment**.
* Hands-on experience with AWS services such as **Amazon EC2, EMR, Glue, S3, RDS, VPC, and IAM.**
* Expertise in using Amazon Web Services such as **EC2 and S3** to analyze and store small data processing. Extensive experience administering the **Hadoop cluster** running on **AWS EMR**.
* Good understanding of the **Snowflake Data Platform**. Experience with Snowflake **Multi-Cluster Warehouses** and **Snow Pipe**, importing data from the local system and AWS S3 Bucket
* Experience working on Snowflake modeling and highly proficient in **data warehousing** techniques for **data cleansing**, **Slowly Changing Dimension (SCD)** and **change data capture (CDC).**
* Good experience on building batch data pipelines using **Azure Databricks Notebooks**.
* Working knowledge of **Object-Oriented Programming (OOP)** principles using **Python**.
* Good understanding of the **Snowflake Data Platform**. Experience with Snowflake **Multi-Cluster Warehouses** and **Snow Pipe**, importing data from the local system and AWS S3 Bucket, in-depth knowledge of **Snowflake Database, Schema** and **Table structures**, and experience with **Snowflake Clone and Time Travel**.
* Good knowledge on implementing the **performance tuning** and various transformations in **snowflake**.
* Extensive knowledge of **Spark Streaming**, **Spark SQL**, and other Spark components such as accumulators, broadcast variables, different cache levels, and Spark optimization approaches.
* Experience with Data Pipelines, ETL and ELT data processes, and converting **Bigdata/Unstructured Data** sets (**JSON**, **Log Data**) to **Structured Data Sets** for Product Analysts and Data Scientists.
* Experience working on **Data Modeling, Data Migration, Design, and ETL Pipeline** for both cloud and on premises. Experienced in **migrating SQL** database to **Azure Data Lake**.
* Experienced in reading Continuous **JSON** data from different source system into **Databricks Delta** and processing the files using **Apache Structured streaming**.
* Good knowledge on implementing **Airflow** to automatically schedule and trigger **data ingestion ETL**/**ELT** pipeline execution.
* Experience in building the **ETL**/**ELT** architecture and source to Target mapping to load data into **Datawarehouse**.
* Expert with **JSON, Avro, Parquet, RC, and ORC Data Formats**, as well as compressions like **Snappy and BZip**.
* Knowledgeable with Reporting tools such as **Power BI and Tableau**.
* Good backend skills, such as designing **SQL Objects** such as **Tables, Stored Procedures, Triggers, Indexes** and **Views** to allow data manipulation and consistency.
* Relationship-driven, results- driven, creative, and able to think outside the box.

**Technical Skills:**

|  |  |
| --- | --- |
| **AWS Cloud Services** | Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambda, Amazon RDS, Amazon Elastic Load Balancing, Amazon SQS, IAM, AWS Cloud Watch, Amazon EBS, AWS Glue |
| **Azure Cloud Services** | Azure Blob Storage, Azure Monitoring, Azure Search, Data Factory, Azure SQL, Azure Analysis Services, Azure Synapse Analytics (DW), Azure Data Lake, Azure Active Directory |
| **Hadoop/Big Data Technologies** | Hadoop, Map Reduce, Pig, Sqoop, Hive, Oozie, Spark, Zookeeper, Flume |
| **Hadoop Distribution** | Horton Works, Cloudera |
| **Dimensional Modelling** | Snowflake and Starschema |
| **Programming & Scripting** | Python, Scala, SQL, Shell Scripting |
| **BI Tools** | Power BI, Tableau |
| **ETL Tools** | Informatica, DataStage |
| **DevOps Tools** | Jenkins, Maven, GIT, SVN |
| **IDEs** | Eclipse, IntelliJ, NetBeans |
| **Operating Systems** | Windows, Linux, Unix and Mac OS |
| **Databases** | Oracle, SQL Server, MySQL, HBase, Mongo DB, Redshift, Snowflake |

**Professional Experience:**

**Client- PenFed Credit Union - McLean, VA**  **02/2021 – Present**

**Senior Data Engineer**

**Responsibilities:**

* Gathered requirements for Analysis, Design, Development, testing, and implementation of business rules.
* Created **data pipelines** and transformed data to a structure that is relevant to the problem by selecting appropriate techniques.
* Performed **ETL operations** for Data cleansing, Filtering, Standardizing, Mapping, and Transforming of Extracted data from multiple sources such **as Azure Data Lake,** and **on-prem SQL DB**.
* Developed Spark applications using **PySpark** and **Spark-SQL** for data transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Worked on predicting the cluster size, monitoring, and troubleshooting the Spark data bricks cluster.
* Designed data warehouses and data lakes on **Oracle, SQL Server**, and high-performance **Teradata** databases.
* Migrated data from on-prem **SQL Database** to **Azure Synapse Analytics** using **Azure Data Factory.**
* Ingested huge volume and variety of data from disparate source systems into **Azure Data Lake** using **Azure Data Factory**.
* Applied various **ADF dataflow transformations** such as Data Conversion, Conditional Split, Derived Column, Lookup, join, Union, Aggregate, pivot, and filter and performed data flow transformation using the data flow activity.
* Created numerous pipelines in **ADF** to get the data from disparate source systems by using different Azure activities like Move & transform, copy, filter, for each, Databricks.
* Worked on developing Application Interface Document for the downstream to create a new interface to transfer and receive the files through **Azure Data Share**.
* Worked on creating pipelines, data flows, and complex data transformations and manipulations using **ADF** and **PySpark** with **Databricks**.
* Worked with IDE tools like **Eclipse** to update and push the data extracted and created workflows from the databases.
* Applied various **ADF dataflow transformations** such as Data Conversion, Conditional Split, Derived Column, Lookup, join, Union, Aggregate, pivot, and filter and performed data flow transformation using the data flow activity.
* Automated and validated the created data-driven workflows extracted from the ADF using **Apache Airflow**.
* Implemented large **Lambda architectures** using Azure Data platform capabilities like **Azure Data Lake, Azure Data Factory, Azure Data Catalog, HDInsight, Azure SQL Server, Azure ML, and Power BI.**
* Scheduled and monitored **data pipelines in ADF** and automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF.
* Orchestrated data pipelines using **Apache Airflow** to interact with services like **Azure Databricks, Azure Data Factory, Azure Data Lake, and Azure Synapse Analytics.**
* Worked on maintaining and tracking the changes using version control tools like **SVN** and **GIT**.
* Worked with building data warehouse structures, and creating facts, dimensions, and aggregate tables, by dimensional modeling, and **Star and Snowflake schemas**.
* Defined **CI/CD** process and support test automation framework in the cloud as part of build engineering team.
* Built **Power BI** reports and dashboards with interactive capabilities.

**Environment:** Azure Synapse Analytics, SQL Database, Azure Data Lake Storage (ADLS), and Azure Data Factory, SQL Database, Azure Synapse Analytics, Azure Data Factory, Teradata, HDFS, Sqoop, Azure Data Lake, Azure Data Factory, ETL, SQL DB, Oracle, SQL Server, Teradata, Azure Data Share, PySpark with Databricks, Apache Airflow

**Client: American Express- Madison, WI 04/ 2018– 12/2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Worked on processing the real-time data analytics using **Spark Streaming, Kafka, and Flume**.
* Configured Spark streaming to get ongoing information from Kafka and store the stream information to **HDFS**.
* Transformed and cleansed the input data extracted from the external data using **Spark**.
* Migrated an existing on-premises application to **AWS** and used **AWS services** like EC2 and S3 for processing and storage of small data sets, and maintained the **Hadoop cluster** using AWS EMR.
* Developed a **Data pipeline** using **Spark, Hive, Impala, and HBase** to analyze customer behavioural data and financial histories in the **Hadoop cluster**.
* Performed transformations using **Spark** and saved the result back to **HDFS** and later to the target database **Snowflake**.
* Designed and developed ETL processes in **AWS Glue** to migrate data from external sources like **S3**, or file formats like **JSON, Parquet and Text Files** into **AWS Redshift**.
* Created **DataStage** jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, and Row Generator.
* Responsible for creating, debugging, scheduling, and monitoring jobs using **Airflow** for ETL batch processing to load into **Snowflake** for analytical processes.
* Built **ETL pipeline** for **data ingestion, data transformation, and data validation** on cloud service AWS, working along with data steward under data compliance.
* Worked on scheduling and validating all jobs with Airflow scripts using **Python** and added different tasks to directed acyclic graph and dependencies between the tasks using **Lambda**.
* Developed **Python scripts** to manage AWS resources from **API calls** using **BOTO3 SDK** and worked with AWS CLI.
* Worked on extracting and filtering the data in data pipelines using **PySpark** and transformed the data pipelines.
* Monitored the servers using **CloudWatch**, and stored and retrieved the data extracted from the workflows.
* Used Spark applications using **Spark-SQL** in Data bricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Worked on writing **Hive and Spark** queries for optimization using window functions and customizing the Hadoop shuffle.
* Integrated **Oozie** with the rest of the Hadoop stack, supporting several types of Hadoop jobs **MapReduce, Hive**, and **Sqoop** and system-specific jobs such as Java programs and Shell scripts.
* Successfully implemented **ETL solutions** between an **OLTP** and **OLAP** database in support of Decision Support Systems with expertise in all phases of SDLC.
* Worked on estimating the cluster size, monitoring, and troubleshooting **Spark Databricks cluster** and scripting using UNIX shell scripting for automating the data load processes.
* Worked on implementing the monitoring solutions using **Docker and Jenkins**.
* Implemented dashboard designs and created worksheets and data visualization dashboards using **Tableau**.

**Environment:** Python, Hive, Spark, AWS EC2, S3, AWS EMR, AWS Glue, HDFS, Spark Streaming, Kafka, MapReduce, Hive, Sqoop, Oozie, Flume, Impala, HBase, JSON, Parquet, Text Files, AWS Redshift, Spark, DataStage, Airflow, Snowflake, ETL pipeline, Lambda, PySpark, Cloud Watch, Spark-SQL, Unix shell, Agile, Tableau, Docker, and Jenkins

**AmeriHealth Administrators Philadelphia, PA 09/2016 - 03/2018**

**Role: Big Data Engineer**

**Responsibilities:**

* Worked on cloud technologies with **Azure** which will use for stages to load data into Snowflake.
* Performed **Data Ingestion** to one or more **Azure Services** (Azure Data Lake, Azure Storage, Azure SQL, and Azure DW) and processed the data in Azure Databricks.
* Implemented medium to large scale BI solutions on Azure using **Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks and NoSQL DB).**
* Migrated on-premise data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake and Stored (ADLS) using **Azure Data Factory (ADF V1/V2)**.
* Implemented the **SnowSQL** and Snow pipe for continuous data load from csv files to snowflake.
* Redesigned the views from SQL server to snowflake by replacing the snowflake functions with **SQL server functions** in Snowflake.
* Implemented the cloning, data retention periods, and the fail-safe mechanism in **Snowflake** to recover the data.
* Loaded the data from **Azure Blob Storage** as an External storage to snowflake which includes continuous loading of data.
* Primarily involved in Data Migration using **SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell**.
* Creating pipelines, data flows and complex data transformations and manipulations using Azure and PySpark with **Databricks**.
* Created numerous pipelines in Azure using **Azure Data Factory** v2 to get the data from disparate source systems by using different activities like Move & Transform, copy, filter for each Databricks etc.
* Created several **Databricks** **Spark jobs** with PySpark to perform several tables to table operations.
* Designed and developed multiple applications in Spark using Python.
* Created and selected the virtual warehouses in **Snowflake** based on the query load performance.
* Designed and developed **Scala workflows** for data pull from cloud-based systems and applying transformations on it.
* Implemented **Apache-Spark code** to read multiple tables from the real-time records and filter the data based on the requirement.
* Used **Spark-Streaming APIs** to perform necessary transformations and actions on the data got from Kafka.
* Worked on analyzing **Hadoop clusters** and different big data analytic tools including **Pig, Hive**.
* Migrated **MapReduce jobs into Spark jobs** and used **SparkSQL** and Data frames API to load structured data into Spark clusters.
* Automated the **SQL Scripts in Snowflake using the Tasks.**
* Wrote **UDF’s and Stored Procedures** to load the data from one phase to another phase in Snowflake.

**Environment:** Azure (Blob Storage, Data Lake, Databricks, SQL Azure, Azure Data Factory), Snowflake, Data Warehousing, Hadoop, Pig, Hive, MapReduce, Spark, Spark SQL, Power BI, Jira, SQL Server, GitLab, SQL, SSIS, PowerShell

**Client: Gateway Technolabs - Hyderabad, 05/2013 - 12/2014**

**Role: Hadoop Engineer**

**Responsibilities:**

* Handled Responsibilities likes managing and monitoring the Hadoop cluster with **Cloudera Manager.**
* Handled massive data sets in parallel across the Hadoop cluster, I have created a **MapReduce** application in Java using the **MapReduce programming framework**.
* Acquired practical knowledge of **Amazon Web Services' (AWS) infrastructure S3 and EC2** are two of Amazon’s cloud services (Amazon EC2).
* Developed Sqoop scripts to import, export, update the data between **HDFS/Hive** and Relational Databases.
* Used Hive Queries to get information stored in **HDFS** and the **Hive Warehouse**.
* Involved in **Directed Acyclic Graph (DAG**) of operations and control flows, defined by Oozie for scheduling tasks to manage **Apache Hadoop workloads.**
* Created **Spark apps in Scala**, packaging them in jars, and running them on a cluster.
* Optimized data processing times by creating Spark core and Spark SQL scripts in **PySpark**.
* Worked on managing and reviewing the Hadoop log files.
* Transformed the data from online transaction processing to online analytical processing, a thorough ETL design was developed.
* Worked on documentation like including a **Source-to-Target Data Mapping Document**, a Set of Unit Test Cases, and a Set of Data Migration Documents.
* Using **SparkSQL and other PySpark libraries,** worked on Data Exploration to analyze patterns and select features.
* Analyzed, discovered, and fixed many Python problems in the two key applications utilized by the company's clients and customer service workers.
* Managed and reviewed **Hadoop Log Files**, deploying and Maintaining Hadoop Cluster
* Responsible for **cluster maintenance**, adding and removing cluster nodes and cluster monitoring.
* Loaded and transformed large amounts of structured, semi-structured, and unstructured data into the **Hadoop system**.
* Used **Spark** to handle streaming data and did sophisticated analytics while processing.
* Contributed to Impala to enable massively parallel **Hive query processing**.
* Used tableau to generate reports, graphs, and charts that offer an overview of the presented data.
* Developed data ingestion pipeline from **HDFS into AWS S3 buckets using Nifi.**

**Environment:** Python,Hadoop, HDFS, MapReduce, HBase, Hive, Sqoop, Oozie, Flume, Nifi, Spark SQL, Spark Context, Amazon Web Services (AWS), S3, and Tableau