**Naveen Rayapati**

**Contact: (913) 335-0072**

**Email: naveen.82393@gmail.com**

**PROFESSIONAL SUMMARY:**

* **8+** years of professional IT experience with over **5** years of experience in ingestion, storage, querying, processing, and analysis of big data systems. Strong experience with programming languages like SQL, Java, Scala, and Python. Good knowledge with architecture of distributed storage systems like **HDFS**, **S3** and parallel computing frameworks like **Spark** and **MapReduce**.
* Strong working knowledge of **Kafka**, **Hive**, **MapReduce**, **HBase**, **Snowflake**, **Sqoop**, **Impala**, **Yarn**.
* Strong experience in utilizing **Spark RDD**, Spark Data Frames, **Spark SQL**, and **Spark Streaming Api’s**.
* Experienced in building and managing distributed data systems on cloud platforms, such as **AWS** and **Azure**, ensuring seamless data processing and storage.
* Good Experience Working with **Azure** cloud services like **ADLS**, HDInsight, **Azure** **Databricks**, Azure Functions, Data Factory, **Synapse** etc.
* Development of **ETL**/**ELT** using Python/ADF to integrate, filter, transform data from multiple source systems.
* Good Experience Working with AWS cloud services like **S3**, **EMR**, **EC2**, Lambda functions, **Redshift**, **Athena**, **Glue** etc.
* Managed batch data from **AWS** **S3** and performed necessary transformation using **Spark** through **EMR**.
* Proficient in designing and optimizing **data** **warehouses**, enabling efficient **data retrieval** and analytics with a focus on data quality and governance.
* Designed and constructed scalable data processing pipelines within **Flink** for efficient stream and batch processing.
* Proven experience handling various file formats such as CSV, TSV, Parquet, ORC, JSON, and AVRO.
* Expertise in writing **DDLs** and **DMLs** scripts in **SQL** and **HQL** for analytics applications in RDBMS.
* Strong **SQL** and **PL-SQL** skills with a focus on working with Oracle databases.
* Knowledge in Full Life Cycle development of **Data Warehousing**.
* Familiarity with **SonarQube** code review, GIT source control, **JENKINS** continuous integration and **JIRA** project management.
* Proficient in designing, developing, and implementing complex data integration workflows using **Informatica** **PowerCenter** to efficiently extract, transform and load data from diverse sources.
* Implemented and optimized Cockroach DB clusters to handle large-scale data processing, ensuring high availability, fault tolerance and data consistency.
* Worked extensively on **Sqoop** for performing both batch loads as well as incremental loads from relational databases.
* Experienced in conducting performance testing of **ETL** workflows to identify bottlenecks and optimize data processing times for large-scale data volumes.
* Created custom dashboards and reports using tools like **Tableau**, **Power** **BI** to visualize SaaS data effectively.
* Served as a subject matter expert data integration, with proficiency in **Informatica**, **Snap Logic**.
* Experienced in setting up and managing **Kubernetes** clusters for container orchestration, ensuring high availability and scalability.
* Worked on implementing various **Airflow** automations for building integrating between **clusters**.
* Proficient understanding of entire Software Development Life Cycle (SDLC), **Agile** Software Development, **Scrum** Methodology, and Test-Driven Development.
* Proficient in designing and implementing **Dimensional Data Models**, including **Star Schema** and **Snowflake Schema**, to facilitate efficient querying and reporting.
* Skilled in implementing data quality checks and **validations** during **ETL** process to identify and rectify data anomalies, ensuring high data integrity.
* Participated in data preparation tasks for **downstream** **machine** **learning** and reporting teams, including data cleansing, event enrichment, data aggregation de-normalization and data preparation.
* Skilled in **data modelling** and database management, implementing **robust** data schemas and ensuring data security and compliance with industry standards.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Big Data | **Spark, Hive, Pig, Map Reduce, Yarn, Kafka, Sqoop, HDFS, Informatica** |
| Languages | **Python, Java, Scala** |
| Scripting | **SQL, JavaScript, CI/CD, Terraform** |
| Azure Components | **Databricks, ADLS, Delta Lake, HDInsight, Synapse Analytics, Azure Functions, ADF, Azure SQL DB** |
| AWS Components | **EC2, EMR, S3, RDS, CloudWatch, Athena, Red Shift, DynamoDB, Lambda** |
| SQL Databases/ Data Warehouse | **Oracle, MySQL, PostgreSQL, Teradata, Snowflake** |
| NO SQL Database | **HBase, DynamoDB, MongoDB** |
| Tools | **Jenkins, JIRA, GitHub, IntelliJ, ServiceNow, VS Code** |
| Analytics Tool | **Tableau, Power Bi** |

**PROFESSIONAL EXPERIENCES:**

**Texas Medicaid & Healthcare Partnership, Austin, TX. July 2022 to Present**

**Sr Data Engineer**

* Involved in development of an **Azure** cloud-based central data lake.
* Hands – on experience **in Azure cloud services, Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Azure Monitoring, Key Vault, Azure Data Lake**.
* Collaborated with **stakeholders** to understand business requirements to ensure data quality and governance.
* Designed and implemented complex data integration pipelines using **Azure Data Factory**, enabling seamless data movement and transformation across diverse data sources.
* Managed and maintained Azure SQL Databases, ensuring data integrity, security, and high availability.
* Developed ETL jobs that need implementation of notebooks in Synapse SQL pool and PySpark pool as well as extraction of data from Oracle and SQL.
* Possessed knowledge of initiating and shutting down runtimes automatically using the Databricks API’s while working with **Azure Databricks** runtimes.
* Developed and implemented **Databricks** notebooks in **Python** using **PySpark** Framework to execute data preparation tasks for machine learning and reporting teams, including data cleansing, event enrichment, data aggregation, and de-normalization.
* Designed and implemented end-to-end ETL pipelines using **Azure Databricks**, extracting data from various sources.
* Used **ADLS** as a **data lake** and ensured that all processed data from **spark** and **hive** jobs was uploaded directly to ADLS.
* Implemented incremental data loading strategies using **SSIS** to optimize performance and minimize resource usage during daily batch processes.
* Integrated error handling and logging mechanisms into **SSIS** packages to capture failed records, exceptions, and package execution status for enhanced monitoring.
* Developed comprehensive **SaaS** reporting solutions to track my key performances metrics and deliver actionable insights.
* Involved in development of **Azure Cloud** platform's data lake, which will enable business teams to conduct data analysis using **Azure Synapse SQL.**
* Working knowledge of **Databricks API** for automating start-up and shutdown of runtimes using **Azure Databricks.**
* Developed advanced analytics models in **Snowflake** for risk assessment, including credit risk, market risk, and operational risk analysis.
* Designed, developed, and maintained complex data integration workflows using **Informatica PowerCenter** to seamlessly integrate financial data from various sources including banking systems, financial markets, and payment gateways.
* Used **Snowflake** to process ETL pipelines and load external data from cloud using stages and workflows.
* Examined data quality issues using **Snow SQL** by building analytical warehouses on **Snowflake**.
* Generating interactive **Power BI** and **Tableau** dashboards.

**Tools Used**: Azure cloud, Azure Data Lake Storage, Hadoop, Hive, Informatica, Python, Spark, Map Reduce, GIT, Jenkins.

**Garmin International, Olathe, KS. Jan 2021 to June 2022**

**Data Engineer**

* Worked on building centralized **Data Lake** on AWS Cloud utilizing primary services like **S3**, **EMR**, **Redshift** and **Athena**.
* Designed and implemented end-to end data solutions using **AWS Glue**, **AWS S3** and **AWS Redshift**, resulting in optimized data processing.
* Worked on migrating datasets and **ETL** workloads with **Scala** from On-prem to **AWS Cloud services**.
* Managed database instances for various relational database engines like MySQL, PostgreSQL, Oracle, SQL Server using **AWS RDS.**
* Extensive experience in utilizing **ETL** Process for designing and building very large-scale data using **Apache** **Spark**
* Migrating data from the local **Teradata** data warehouse to **AWS S3** data lakes.
* Designed and implemented **DynamoDB** database schemas to efficiently store and retrieve structured and semi-structured data.
* Integrated Electronic Health Records (EHR) data into **Snowflake**, enabling comprehensive analysis and reporting for healthcare providers.
* Integrated **AWS RDS** into ETL workflows to extract data from various sources, transform it to fit business requirements, and load it into target database.
* Designed and optimized data models in **Snowflake** to support advanced analytics, including predictive modeling, patient outcome analysis, and population health management.
* Orchestrated containerized applications using **Kubernetes**, enabling efficient scaling and deployment.
* Worked extensively on building and automating **data ingestion pipelines** and moving terabytes.

of data from existing **data warehouses** to cloud.

* Worked extensively on fine tuning **spark** applications and providing production support to various pipelines running in production.
* Designed and implemented automated data integration processes to streamline **SaaS** **reporting**, reducing manual data entry and ensuring data accuracy.
* Developed Python code to gather data from **HBase** and designs solution to implement using **PySpark**.
* Developed and optimized **Python** based **ETL** **pipelines** in both legacy and distributed environments.
* Designed end-to-end data pipelines with **Snap Logic** to extract, transform and load clinical data from multiple health systems into a centralized data repository for analysis and reporting.
* Developed Spark with Python based pipelines using spark data frame operations to load data to **EDL** using **EMR** for jobs execution & **AWS S3** as storage layer.
* Worked on a full spectrum of data engineering pipelines: data ingestion, data transformations and data analysis/consumption with **Python**.
* Extracted data from **AWS Aurora Databases** for big data processing.
* Used **Scala** data pipelines to perform transformations on **EMR** clusters and loading the transformed data into S3 and from **S3** into **redshift**.
* Worked on creating **Kafka** producers using **Kafka** **Java** **Producer** **API** for connecting to external Rest live stream application and producing messages to **Kafka** **topic**.

**Tools Used**: AWS S3, EMR, Redshift, Aurora, Snap Logic, Athena, Glue, Kubernetes, Spark, Scala, Python, Java, Hive, Kafka

**Target,** **Brooklyn Park, MN. Apr 2019 – Dec 2020**

**Data Engineer**

* Developed series of **data ingestion** jobs for collecting data from multiple channels and external applications in **Scala**.
* Worked on both batch and streaming ingestion of data.
* Built **Python** based **Data pipelines** from multiple data sources by performing necessary **ETL** Tasks.
* Imported clickstream log data from **FTP** Servers and performed various data transformations using Spark Data frame API and **Spark-SQL** API’s.
* Implemented **Java** based **Kafka** Producer applications for streaming messages to Kafka topics.
* Built **Spark** Streaming applications for consuming messages and writing to **HBase**.
* Worked on troubleshooting and optimizing **Spark** Applications.
* Worked on ingesting data from SQL-server to S3 using **Sqoop** with in **AWS EMR**.
* Migrated Map-reduce jobs to Spark applications built on Scala and integrated with **Apache Phoenix** and **HBase**.
* Developed daily and monthly **ETL** processes which were automated using custom **UNIX shell scripts** &
* **Python**.
* Worked on building **ETL** pipelines using **Python** scripting, Pandas Data Frames.
* Integrated **Kubernetes** into CI/CD pipelines, automating application deployments and ensuring continuous delivery.
* Involved in loading and transforming large sets of data and analyzing them using **Hive** Scripts.
* Implemented **SQL** queries on **AWS** with platforms like **Athena** and **Redshift**.
* Optimized resource allocation in **Kubernetes** clusters, minimizing infrastructure costs while maintaining performance.
* Have experience in Querying in **AWS Athena** where alerts are coming from **S3** buckets and finding difference in time interval between Clusters of **Kafka** and **Kinesis**.
* Loaded portion of processed data into **Redshift** tables and automated the process.

**Tools Used**: Hadoop, Spark, Scala, Hive, Sqoop, Python, Oozie, Kafka, AWS EMR, Redshift, S3, Kinesis, Kubernetes, Spark Streaming, Athena, HBase, JIRA, Shell Scripting, Maven, Git

**United Health Group (UHG), Bellevue, WA. Nov 2018 – Mar 2019**

**Big Data Developer**

* Involved in importing and exporting data between **Hadoop Data Lake** and Relational Systems like **Oracle**, **MySQL** using **Sqoop**.
* Involved in developing spark applications to perform **ETL** kinds of operations on data.
* Modified existing **MapReduce** jobs to Spark transformations and actions by utilizing **Spark RDDs**, **Data Frames** and **Spark SQL API’s.**
* Handled batch data from **AWS S3** and performed necessary transformation using Spark through

**EMR**.

* Designed a data analysis pipeline in **Python**, using **Amazon Web Services** such as **S3**, **EC2** and **Elastic** **MapReduce.**
* Utilized Hive partitioning, **Bucketing** and performed various kinds of joins on **Hive** tables.
* Involved in creating **Hive** external tables to perform **ETL** on data that is produced on a daily basis.
* Leveraged scripting languages and automation tools to streamline routine **Kubernetes** tasks, improving efficiency.
* Validated data being ingested into **Hive** for further filtering and cleansing.
* Developed **Sqoop** jobs for performing incremental loads from **RDBMS** into **HDFS** and further applied Spark transformations.
* Loaded data into **hive** tables from **spark** and used **Parquet** columnar format.
* Identified and pooled valuable data from various sources to extract, process and analyze data for feature engineering and model selection.
* Created **Oozie** workflows to automate and productionize the **data pipelines.**
* Migrating **Map Reduce** code into Spark transformations using **Spark** and **Scala**.
* Collecting and aggregating large amounts of log data using **Apache Flume** and staging data in **HDFS** for further analysis.
* Designed and documented operational problems by following standards and procedures using **JIRA.**

**Tools Used**: Hadoop, Hive, Impala, Python, Oracle, Spark, Kubernetes, Pig, Sqoop, Oozie, Map Reduce, GIT

**Hive loop Technologies, Bangalore, India. Dec 2016 – Sep 2018**

**Hadoop/Java Developer**

* Built custom Input adapters for ingesting gigabytes of behavioral event logs from external servers such as **FTP** **server** and **S3** buckets on daily basis.
* Created **Sqoop** scripts to import/export user profile and other lookup data from **RDBMS** to **S3** data store.
* Developed **ETL** programs using Informatica to implement business requirements.
* Communicated with business customers to discuss issues and requirements.
* Created shell scripts to fine tune ETL flow of Informatica workflows.
* Developed various **spark** applications using **Scala** to perform cleansing, transformation and enrichment of these click stream data.
* Involved in data cleansing, event **enrichment**, **data** **aggregation**, **de-normalization,** and **data.**

**preparation** needed for **machine learning** and reporting.

* Utilized **Spark RDD**, **Data Frames** and **Spark SQL API** to implement batch processing of jobs.
* Troubleshooting **Spark** applications for improved error tolerance and reliability.
* Fine-tuning **spark** applications/jobs to improve efficiency and overall processing time for **pipelines**.
* Created **Kafka** producer API to send live stream JSON data into various Kafka topics.
* Developed **Spark**-Streaming applications to consume data from **Kafka** topics and to insert processed streams to **HBase**.

**Tools Used**: AWS EMR, Python, Spark, Hive, HDFS, Sqoop, Kafka, Oozie, HBase, Scala, Java, Informatica