**Sr. Data Engineer**

**Naimisha Doddi**

**Email**: [naimisha61@gmail.com](mailto:naimisha61@gmail.com) **Contact number:** 9784968676

**Professional Summary:**

* **10 years** of professional experience in information technology with an expert hand in the areas of **Big Data, Hadoop Spark, Hive, Impala, Sqoop, Flume, Kafka, SQL tuning, ETL development, report development, SAS, database development, data modeling** and strong knowledge of **oracle database architecture**.
* Experience **in Big Data analytics, Data manipulation, using Hadoop Eco system tools Map - Reduce, HDFS, Yarn/MRv2, Pig, Hive, HDFS, HBase, Spark, Kafka, Flume, Sqoop, Flume, Oozie, Avro, Sqoop, GCP, AWS, Spring Boot, Spark integration with Cassandra, Avro, Solr and Zookeeper**.
* Strong experience in migrating other databases to **Snowflake**.
* Managing Database, **AWS Data Platform** services (**AWS Data Lake (ADLS**), **Data Factory (ADF), Data Lake Analytics**, **Stream Analytics, AWS SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.**
* Extensive experience in Text Analytics, generating data visualizations using R**, Python** and creating dashboards using tools like **Tableau, Power BI**.
* Proficient in ArcGIS Pro and spatial data handling. Seeking opportunities to apply data engineering skills and ArcGIS Pro experience to contribute to innovative projects in a dynamic environment.
* Designed and implemented complex ETL processes to transform and load data from various sources into Snowflake, ensuring data consistency and accuracy.
* 2 years of experience in data engineering with a focus on spatial data management and analysis using **ArcGIS** **Pro**.
* Experienced in building Automation Regressing Scripts for validation of **ETL** process between multiple databases like **Oracle, SQL Server, Hive**, and **MongoDB** using **Python**.
* Conducted training sessions for the team to enhance Snowflake proficiency.
* Utilized **Kubernetes** and **Docker** for the runtime environment for the **CI/CD** system to build, test, and deploy. Experience in working on creating and running **Docker** images with multiple **micro services.**
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like **Map Reduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.**
* Proficient with **Spark Core, Spark SQL, Spark MLlib, Spark GraphX** and **Spark Streaming** for processing and transforming complex data using in-memory computing capabilities written in **Scala.** Worked with Spark to improve efficiency of existing algorithms using **Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.**
* Skilled in implementing data governance policies, data quality standards, and data security measures.
* Experience in designing, building, and implementing complete Hadoop ecosystem comprising of **Map Reduce, HDFS, Hive, Impala, Pig, Sqoop, Oozie, HBase, MongoDB,** and **Spark.**
* Experience with Client-Server application development using **Oracle PL/SQL, SQL PLUS, SQL Developer, TOAD,** and **SQL LOADER.**
* **Used AWS EMR Spark cluster and Cloud Dataflow on GCP** to compare the efficiency **of a POC on a developed pipeline**
* Working experience in **migrating** several other databases to Snowflake
* Strong experience with architecting highly per formant databases using **PostgreSQL, Post GIS, MySQL** and **Cassandra.**
* Experience with AWS Cloud formation and launch configurations to automate repeatable provisioning of AWS resources for application with strong hands on experience with AWS solutions using EC2, S3, RDS, and Elastic Load Balancer, Auto Scaling groups, Lambda, API Gateway and EBS services.
* Extensive experience in loading and analyzing large datasets with **Hadoop framework (MapReduce, HDFS, PIG, HIVE, Flume and Sqoop)**
* Having extensive experience in **Microsoft AWS Cloud Computing, GCP** **and SQL BI Technologies**.
* Hands-on experience in **Azure Cloud Services (PaaS & SaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake**.
* Good experience in tracking and logging end to end software applications build using **Azure DevOps**.
* Experienced with **Docker** and **Kubernetes** on multiple cloud providers, from helping developers build and containerize their application **(CI/CD)** to deploying either on public or private cloud.
* Successfully migrated ETL applications from Informatica PowerCenter to Informatica Intelligent Cloud Services (IICS).
* Developed complex mappings, transformations, and workflows in Informatica PowerCenter and IICS for large-scale data integration projects.
* Developed ETL process using Informatica Power center & IDMC
* Strong experience in core **Java, Scala, SQL, PL/SQL** and **Restful web services.**
* Good experience working on analysis tool like **Tableau** for **regression analysis, pie charts, and bar graphs.**
* Experienced on implementation of a log producer in **Scala** that watches for application logs, transform incremental log and sends them to a **Kafka** and **Zookeeper** based log collection platform.
* Have hands-on experience working with source control tools such as Clear Case, SVN, GIT and Surround SCM. Strong expertise in developing multi-tier web application projects using Spring Boot, NodeJS server and deploying the applications on to the AWS cloud with Devops tools.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data/Hadoop Technologies** | Map Reduce, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark, ,Pig, Hive, HBase, Flume, Yarn, Oozie, Zookeeper, Hue, Ambari Server,Data Governance, AWS |
| **Languages** | Java, Scala, Python (NumPy, SciPy, Pandas, Genism, Keras), Shell Scripting, Pyspark |
| **NO SQL Databases** | Cassandra, HBase, MongoDB, Maria DB |
| **Web Design Tools** | HTML, CSS, JavaScript, JSP, jQuery, XML |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Azure Data bricks, Eclipse, NetBeans, AWS. |
| **Public Cloud** | EC2, IAM, S3, Auto scaling, Cloud Watch, Route53, EMR, Redshift, GCP |
| **Development Methodologies** | Agile/Scrum, UML, Design Patterns, Waterfall |
| **Build Tools** | Jenkins, Toad, SQL Loader, PostgreSQL, Talend, Maven, ANT, RTC, RSA, Control-M, Oozie, Hue, SOAP UI |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio/Outlook), SSRS, Cognos |
| **Databases** | Microsoft SQL Server, MySQL, Oracle 11g, 12c, DB2, Teradata, Netezza |
| **Operating Systems** | All versions of Windows, UNIX, LINUX, Macintosh HD, Sun Solaris |

**PROFESSIONAL EXPERIENCE:**

**Client: Elevance Health June 2023 – till date**

**Role: Senior Data Engineer**

**Location:GA**

**Responsibilities:**

* Import the data from different sources like **HDFS/HBase** into **Spark RDD** and developed a data pipeline using **Kafka** and **Storm** to store data into **HDFS**.
* Designed and developed **AWS** Cloud relational servers and databases analyzing current and future business requirements.
* Over 4 of experience as a **Data Engineer** with a specialization in geospatial data and **ArcMap**.
* Worked on migration of data from on Prem SQL server to **Cloud database (AWS)**.
* Good work experience in **RDBMS** such as **MySQL** and SQL server and NoSQL Database like Mongo DB, Cassandra.
* Created tabular models on **AWS** analysis services for meeting business-reporting requirements.
* Led the migration of on-premises data warehouses to **Snowflake**, resulting in a 30% reduction in operational costs and improved query performance.
* Created data models, schemas, and views in **Snowflake** to support business intelligence initiatives and decision-making.
* Used Python, PySpark, shell script, oracle scheduler, Luigi, Oracle Pl SQL etc.
* Transferred Data from Data Lake comprised of **AWS** buckets where the raw events are stored into DataMart which is loaded via Spark streaming pipelines.
* Responsible for continuous monitoring and managing **Elastic MapReduce (EMR)** cluster through **AWS** console.
* Wrote and **PySpark** codes in **Scala** to run a sorting application on the data stored on **AWS**.
* Deployed the application jar files into **AWS** instances.
* Optimized **ETL** workflows by leveraging Informatica's pushdown optimization and parallel processing features, leading to significant performance improvements.
* Strong experience in working with NoSQL databases like **MongoDB**
* Utilized **ArcMap** and **GIS** tools to process, manage, and analyze geographic data.
* Successfully led the migration of a legacy data warehouse to **Snowflake**, including data modeling and **ETL** processes. Experienced in utilizing **ArcMap** and **GIS** tools to analyze geographic data and spatial information
* Led the end-to-end implementation of **ETL** processes using Informatica PowerCenter.
* Working with **NoSQL** databases like **MongoDB** to store and retrieve data.
* Used **AWS** services like **ADLS** and **Synapse** **analytics** for small data sets.
* Developed **PYSpark** and **PySpark-SQL** code to process the data in Apache Spark on **AWS** to perform the necessary transformations based on the **STMs** developed.
* Worked on Version control systems like **Subversion**, **AWS DevOps, GIT, CVS**.
* Developed business intelligence solutions using **SQL** Server data tools to load data to **SQL & AWS Cloud databases**.
* Designed and implemented **Unix-based** solutions for data integration in collaboration with cross-functional teams, ensuring data consistency across various sources and formats.
* Expert in writing **SQL** queries and using Hibernate framework with Spring **ORM** in interaction with the **RDBMS** and familiar with Non-Relational Database (NoSQL) like **MongoDB** and **Cassandra**.
* Utilized **ArcGIS** Pro for processing, managing, and analyzing spatial data for diverse projects.
* Data governance involves defining data quality standards and monitoring compliance with these standards.
* Proficient in designing, deploying, and managing data processing pipelines on **Google Kubernetes Engine (GKE)** to ensure scalable and containerized data applications.
* Used Airflow for scheduling the **Hive**, **Spark, Pyspark** ,and **Map** **Reduce** jobs.
* Developed Oozie work processes for planning and arranging the **ETL** cycle. Associated with composing **Python** scripts to computerize the way towards extricating weblogs utilizing **Airflow** **DAGs**.
* Transformed date related data into application compatible format by developing **Apache Pig UDFs**.
* Configure Zookeeper to coordinate and support **Kafka**, **PySpark**, **Spark** **Streaming**, **HBase** and **HDFS**.
* Programmer analysts with expertise Tableau Servers in **ETL**, **Teradata** and other **EDW** data integrations and developments.
* Assisted in the analysis of spatial data using **ArcGIS** Pro for this project.
* Dedicated and highly skilled Data Engineer with a proven track record in designing, developing, and maintaining data pipelines using **Apache** **Flink**.
* Demonstrated proficiency using **Unix scripting** to parse and manipulate both structured and unstructured data formats, facilitating seamless data transformations and analysis.
* A proven troubleshooter who can identify and fix problems with **Unix scripts** and data pipelines while improving performance and reducing downtime.
* Contributed to the design and development of real-time fraud detection system using **Apache** **Flink**, reducing fraudulent transactions by 25%.
* Integrated **Kafka** with **Spark** Streaming for real time data processing.
* Worked on Implementation of a log producer in **Scala** that watches for application logs transform incremental log and sends them to a **Kafka** and **Zookeeper** based log collection platform
* Worked on Oracle Databases, Redshift and Snowflakes.
* Documented the requirements including the available code, which should be implemented using **Spark, Hive, HDFS**, **HBase** and **Elastic** **Search**.
* Transformed and aggregated data for analysis by implementing work flow management of **Sqoop**, **Hive** and **Pig** **scripts**.
* Involved in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, **HBase, Oozie, Zookeeper, SQOOP, flume, Spark, Impala, and Cassandra** with Horton work Distribution.
* Developed **Kafka** consumer **API** in Scala for consuming data from **Kafka** topics.
* Designed and implemented real-time data processing pipelines using **Apache Flink**, ensuring low-latency data ingestion and processing for critical business applications.
* Designed and implemented **Sqoop** for the incremental job to read data from **DB2** and load to Hive tables and connected to Tableau for generating interactive reports using **Hive** **server2**.
* Used **Sqoop** to channel data from different sources of **HDFS** and **RDBMS**.
* Developed **Scala scripts** using both **Data frames/SQL** and **RDD/Map** Reduce in **Spark** for Data Aggregation, queries and writing data back into **OLTP** system through **SQOOP**.
* Integrated Oozie with Hue and scheduled workflows for multiple **Hive, Pig and Spark** Jobs.
* Developed Spark applications using **Pyspark** and **Spark**-**SQL** for data extraction, transformation and aggregation from multiple file formats.
* building extensive **Hive & SPL quires** to understand behavior across the customer life cycle.
* Developed data connectors to integrate **Apache Flink** with various data sources and sinks, including databases, cloud storage, and third-party **APIs**.
* Worked on data modeling and schema design to ensure data consistency and quality, and implemented data validation checks to identify and rectify data anomalies.
* Created and Managed Splunk,**DB2** connect identities, connections inputs, outputs lookups and access controls.
* Created dashboards, reports and alerts for real time monitoring in **Splunk. Tableau** and Jasper soft.

**Environment**: **HDFS, Hive, Sqoop, Apache nifi, Amazon Web Services (AWS), EMR, Scala, Python, GIT,Jenkins,Hadoop, Oracle, Spark-SQL, PySpark, Kafka, SQL, Oozie, Cosmos, Sqoop, MS Access, HBase, Snowflake, Airflow**, **Tableau.**

**Client: Altice USA Aug 2022 – April 2023**

**Role: Data Engineer**

**Location: NY**

**Responsibilities:**

* Data Ingest from **Sqoop** & **flume** from **Oracle** database.
* Wrote the **Pig UDF** in for converting Date format and time stamp formats from the unstructured files to required date formats and processed the same.
* Hands on Abinitio **ETL**, Data Mapping, Transformation and Loading in complex and high-volume environment.
* Analyzed and managed system logs with **Splunk** and **syslog**.
* Assisted the Product Owner in creating a Proof of Concept by analyzing months of syslog data in the proposed **Splunk** platform in terms of reporting, monitoring and alerts.
* Imported/exported data into **HDFS** and **Hive** using **Sqoop** and **Kafka**.
* Experienced in developing **Map Reduce** programs using **Apache Hadoop** for working with **Big Data.**
* Validate Scoop jobs, Shell scripts & perform data validation to check if data is loaded correctly without any discrepancy. Perform migration and testing of static data and transaction data from one core system to another.
* For data extraction, transformation, and loading (**ETL**) tasks, developed and maintained intricate Unix shell scripts to guarantee accurate and efficient data flow
* Actively involved setting up production support environment in **AWS** using **Kubernetes**.
* Migrated on premise applications to **AWS** cloud using **AWS** services such as **EC2, S3, RDS, Lambda, and Dynamo DB.**
* Proficient in deploying and managing data processing pipelines on **Google Kubernetes Engine (GKE)** for scalable and containerized data applications.
* Frameworks for distributed data processing, including **Apache** **Spark** and **Apache** **Flink**, have been successfully installed on **GKE**, allowing for effective parallel processing of large datasets.
* Applied Apache Kafka to transform live streaming with the batch processing to generate reports.
* Extensively using open source languages **Perl, Python, Scala and Java**.
* Writing scripts for creating, truncating, dropping, altering HBase tables to store the data after execution of MapReduce job and to use that for later analytics.
* Experience in using **AWS** Elastic Beanstalk and CloudFormation for deploying and managing infrastructure on **AWS**.
* **Unix scripts** have error handling and logging mechanisms in place to find and fix data anomalies and processing errors, ensuring the quality and integrity of the data.
* Create Self Service reporting in **Azure** **Data Lake Store Gen2** using an **ELT** approach.
* Developed data warehouse model in snowflake for over 100 datasets using where cape.
* Experience in **Agile** Methodologies, Scrum stories and sprints experience in a **Python** based environment, along with data analytics and data wrangling.
* Communicate workarounds to be followed to L2 and L3 till the issue/work order gets complete/resolve. Raising issues to development team and working with them closely for permanent fixes through calls.
* Developed spring batch applications and used the **AWS** S3 API to store and retrieve the files.
* Developed the applications using **AWS** security best practices (e.g., not using secret and access keys in the code, instead using IAM roles).
* AWS kinesis streams, **AWS** step functions pipelines AWS kinesis streams, Google tensor flow, **AWS** step functions (**AWS, EKS**)
* Performance tune up **Phoenix/HBase, Hive queries and Spark**.
* The custom File System plugin allows Hadoop Map **Reduce programs, HBase, Pig and Hive** to work unmodified and access files directly.
* Writing **PySpark** and **Spark** **SQL** transformation in **Azure** Data bricks to perform complex transformations for business rule implementation
* Design and develop **hive, HBase** data structure and Oozie workflow.
* Extending Hive and Pig core functionality by writing custom **UDFs, UDTF and UDAFs**.
* Build and maintain the environment on **Azure IAAS, PAAS**.
* **GKE** was used to containerize data processing applications, and the microservices architecture was used for increased scalability and flexibility.
* Implement Continuous integration/continuous development best practice using **Azure** DevOps, ensuring code versioning
* Contributed to the design and implementation of cloud data management warehouses using platforms such as **Snowflake**, **Redshift**, or **BigQuery** with **Informatica** **Cloud**.
* Collaborated with business stakeholders to understand data managements and implemented Informatica solutions that aligned with business objectives.
* Architect & implement 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, NoSQL DB).**
* Wrote **Azure** PowerShell scripts to copy or move data from local file system to HDFS Blob storage.
* Managing Database, **Azure** Data Platform services (**Azure** **Data Lake(ADLS), Data Factory(ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB**), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.
* Developed **ETL** jobs using **Spark**-**Scala** to migrate data from **Oracle** to new hive tables.
* Experienced in working with different scripting technologies like Python, Unix shell scripts.
* Developed Spark code using Scala and **Spark-SQL/Streaming** for faster testing and processing of data.
* Develop java-spring based middleware component services to fetch data from HBase using **Phoenix SQL** layer for various **WEB UI** use cases.
* Established data management processes using Informatica Metadata Manager to track and govern data assets in the cloud
* Developed Spark code using Scala and **Spark-SQL/Streaming** for faster testing and processing of data.

**Environment:** **Sqoop, Hive, Azure, Kafka, Snowflake, Python, Oracle, Hadoop, Agile Scrum, Pig, Spark, Scala, Azure Data Bricks, Azure Data Lake**

**Client: Fresenius Medical care June 2015 – July 2022**

**Role: Big Data Engineer**

**Location: MA**

**Responsibilities:**

* Worked in **Agile** environment, and used rally tool to maintain the user stories and tasks.
* Implemented **Apache Sentry** to restrict the access on the hive tables on a group level.
* Designed and implemented by configuring Topics in new **Kafka** cluster in all environment.
* Created multiple dashboards in **tableau** for multiple business needs.
* Implemented Partitioning, Dynamic Partitions and Buckets in **HIVE** for efficient data access.
* Designed **SSIS Packages** to extract, transfer, load (**ETL**) existing data into **SQL Server** from different environments for the **SSAS cubes (OLAP)**
* Extract Transform and Load data from sources systems to **Azure Data Storage** services using a combination of **Azure Data Factory**, **T-SQL**, **Spark SQL**, **Azure Data Lake** analytics.
* Designed & implemented database solutions in **Azure SQL Data Warehouse**, **Azure SQL**.
* Implemented Composite server for the data virtualization needs and created multiples views for restricted data access using a **REST API**.
* Vertical and horizontal scaling strategies were used to optimize the use of resources on **GKE clusters**, effectively controlling infrastructure costs while maintaining performance.
* Exported the analyzed data to the relational databases using **Sqoop** for visualization and to generate reports for the BI team Using **Tableau**.
* Migrated **Map reduce** jobs to **Spark jobs** to achieve better performance.
* Involved in converting **Map Reduce** programs into **Spark** transformations using **Spark RDD's** using **Scala and Python.**
* Worked cross-functionally with software engineers and data scientists to design customized **GKE** solutions that satisfied particular data processing specifications, ensuring peak performance.
* Developed **Apache Spark** applications by using **spark** for data processing from various streaming sources.
* Developed data pipeline using **Spark**, **Hive**, **Pig**, **python**, **Impala**, and **HBase** to ingest customer
* Involved in converting **Hive/SQL** queries into **Spark** transformations using **Spark RDDs**, **Python** and **Scala**.
* Queried and analyzed data from **Cassandra** for quick searching, sorting and grouping through **CQL**.
* Joined various tables in **Cassandra** using **spark** and **Scala** and ran analytics on top of them.
* **GKE** configurations, deployment processes, and troubleshooting procedures were kept in depth documented.
* Applied **Spark** advanced procedures like text analytics and processing using the in-memory processing.
* Implemented Apache Drill on **Hadoop** to join data from SQL and No SQL databases and store it in **Hadoop**.
* Brought data from various sources in to Hadoop and Cassandra using **Kafka**.
* **SQL Server reporting services** (**SSRS**). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports & custom reports.
* Created **Kubernetes** Deployments and Services to manage and expose containers, ensuring high availability and load balancing.
* Designing and Developing **Oracle PL/SQL** and **Shell Scripts**, Data Import/Export, Data Conversions and Data Cleansing.

**Environment:** **Hadoop, Map Reduce, HDFS, Hive, Pig, Impala, Kafka, Cassandra, Spark, Scala, Python, SQL, Zookeeper, Sqoop, Azure**

**Client: Cyber mate InfoTech Ltd Jan 2012 – Jul 2013 Location:Hyderabad, India**

**Role: Data Analyst**

**Responsibilities:**

* Coordinated and tracked project **scope**, **cost**, and schedule.
* Led weekly Debrief sessions to evaluate previous week's team "**Accomplishments**" and "**Open Items**" on **JIRA**.
* Translated business user concepts and ideas into comprehensive business requirements and design documents.
* Drafted Requirements **Traceability Matrix** to ensure the requirements are refined in concise and accurate manner.
* Produced the **RTM**, **BRD** deliverables responsible for the business analyst in the project analysis phase.
* Responsible for writing complex **SQL** for extracting data from the heterogeneous systems using **SQL** **objects** and utilizing these objects as per the need for the business.
* Analyzed set behavior and contribution to business performance, critical business metrics tracking **underlying** business trends using **Business** **Objects**.
* Performed data analysis and data profiling using complex **SQL** on various sources systems including **Oracle**.
* Creating and organizing User stories in **Product** **backlog**.
* Coordinating and **tracking** **project** status to ensure that requirements are met.

**Environment: SQL, PL/SQL, Teradata14, Tableau, T-SQL, Excel.**