**SANDEEP MANAM Email:**

**PROFESSIONAL SUMMARY:**

* **12 years** of professional IT experience and over 5 years of **Big Data** Ecosystem experience in ingestion, storage, querying, processing, analysis of big data/ Databricks/cloud Technologies **(AWS, Azure)**
* Hands-on experience **in Azure Cloud Services (PaaS & IaaS), Azure Databricks, Azure Synapse Analytics, Azure Cosmos DB, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure HDInsight, Key Vault, Azure Data Lake for data ingestion, ETL process, data integration, data migration, AI** solutions.
* Ingested data into **Azure Blob** storage and processed the data using **Databricks**. Involved in writing Spark scripts and **UDF's** to perform transformations on large dataset.
* Experience working with **Azure Blob** and **Data Lake storage** and loading data into **Azure SQL Synapse** analytics.
* Experience in building, maintaining multiple **Hadoop clusters** of different sizes and configuration.
* Created Databricks notebooks to streamline and curate the data for various business use cases and mounted blob storage on Databricks.
* Experience in building data pipelines, computing large volumes of data using **Azure Data factory**.
* Developed Python scripts to do file validations in Databricks and automated the process using **ADF**.
* In-depth knowledge of**Hadoop** and **Spark**, experience with data mining and stream processing technologies **(Kafka, Spark Streaming)**
* Expertise in Big Data architecture like **Hadoop (Azure, Hortonworks, Cloudera)** distributed system, **MongoDB, NoSQL and HDFS**, parallel processing - MapReduce framework.
* Development of Spark-based application to load streaming data with low latency, using **Kafka** and **Pyspark** programming.
* Extensive hands-on experience tuning **spark Jobs**. Experienced in working with structured data using **HiveQL** and optimizing **Hive queries**.
* Hands on experience on Hadoop /Big Data related technology experience in Storage, Querying, Processing and analysis of data.
* Experience in development of Big Data projects using **Hadoop, Hive, HDP, Pig**, **Flume**, and **MapReduce** open-source tools.
* Experience in installation, configuration, supporting and managing **Hadoop clusters**.
* Experience in working with **MapReduce** programs using **Apache Hadoop** for working with **Big Data.**
* Experience in developing, support and maintenance for the **ETL (Extract, Transform and Load)** processes using Talend Integration Suite.
* Experience in installation, configuration, supporting and monitoring Hadoop clusters using Apache, Cloudera distributions and **AWS.**
* Strong hands-experience with AWS services, including **EMR, S3, EC2, Lambda, Glue, Redshift,Athena,DyanmoDB.**
* Excellent working experience in **Scrum / Agile framework** and **Waterfall** project execution methodologies.
* Hands on experience in Hadoop ecosystem including **Spark, Kafka, HBase, Scala, Pig, Hive, Sqoop, Oozie, Flume**, big data technologies.
* Worked on Sparks, Spark Streaming and using **CoreSparkAPI** to explore Spark features to build data pipelines.
* Experienced in working with different scripting technologies like **Python, UNIX** shell scripts.
* Good Knowledge in **Amazon Web Service (AWS)** concepts like **EMR** and **EC2** web services successfully loaded files to **HDFS** from **Oracle**, **SQL Server**, **Teradata** and **Netezza** using **Sqoop.**
* Installed and configured **Apache airflow** for workflow management and created workflows in python.
* Developed python code for different tasks, dependencies, **SLA watcher** and **time sensor** for each job for workflow management and automation using **Airflow tool**.
* I have experience in database design, entity relationships and database analysis, programming **SQL, stored procedures PL/SQL,** packages, and triggers in Oracle.
* Experience in working with different data sources like **Flat files, XML files** and **Databases**.
* Hands on experience in working with Continuous Integration and Deployment **(CI/CD)**
* Strong communication skills, analytic skills, good team player and quick learner, organized and self-motivated.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, Scala, Java |
| **Big Data Technologies** | PySpark, ScalaSpark, YARN, Hive, Sqoop, Spark SQL, HDFS, Map Reduce, NiFi, Kafka Streaming, Spark Streaming |
| **IDE & Tools** | Eclipse, IntelliJ, Oracle (PL/SQL), Putty |
| **Hadoop Platform** | Hortonworks, Cloudera, AWS EMR, EC2 |
| **Cloud services** | Azure(ADF,AKS,AzureAnalytics,HDInsight’s,ADL,Synapse), Aws(S3,EMR,Glue,Redshift,lambda,Athena) |
| **Scripting Languages** | HTML, CSS, JSON, UNIX, Java Script |
| **Database** | Oracle, MySQL, Teradata, MS SQL, Snowflake, DynamoDB |
| **Schedulers** | Airflow, Oozie |

**PROFESSIONAL EXPERIENCE:**

**MasterControl, Salt Lake City, UT Jan 2022 - Present**

**Azure Data Engineer**

**Description:**

Data Engineering team works on building data lake using Bigdata to pull any type of customer information. The challenge is to make all customer data easily accessed for the marketing group so that when they pull groups of data, they are capturing the correct audience through Birst.

**Responsibilities:**

* Worked on end-to-end **ETL** (Extract, Transform, Load) processes using Informatica, ensuring seamless data integration and quality.
* Designed and implemented data solutions in **Microsoft Azure**, including **Azure Data Warehouse** (Azure DW), **Azure Data** **Factory** (ADF), and **Azure Databricks**.
* Developed and maintained data engineering workflows using **Python, SQL**, and Bash scripting for automation and data manipulation.
* Managed and optimized Hadoop ecosystems for large-scale data processing, including **Hadoop Distributed File System** (HDFS), **Spark**, **Hive**, and **Pig**.
* Utilized **Sqoop** and **PySpark** to efficiently transfer and process data between **Hadoop** and other data storage systems.
* Programmed in **Scala** to build robust data processing applications, enhancing data transformation capabilities.
* Administered and optimized **PostgreSQL** and **Cassandra** databases, ensuring data integrity and availability.
* Orchestrated Azure resource provisioning and management through **ARM** templates for **Infrastructure as Code (IaC**).
* Processed and transformed structured and unstructured data in **XML** and **JSON** formats, adapting it for various analytical needs.
* Containerized applications and services using **Docker** and orchestrated them with **Kubernetes** for scalability and portability.
* Managed version control and code repositories using **Bitbucket,** ensuring collaboration and code quality.
* Implemented **CI/CD** pipelines in **Azure DevOps**, automating deployment and ensuring continuous integration.
* Utilized **ELK** **(Elasticsearch, Logstash, Kibana)** stack for log analysis and monitoring data pipelines.
* Experience working with different file formats Parquet, Avro, ORC, and JSON.
* Experienced in handling large datasets using Spark and implementing effective and efficient optimizing techniques.
* like Joins, transformations, and actions.
* Tracked project tasks and progress using **JIRA**, participating in **Agile** methodologies to meet project milestones.
* Created interactive data visualizations and dashboards using **Looker**, providing actionable insights to stakeholders.

**Environment**: ETL, Informatica, Azure, Azure DW, ADF, Azure Databricks, Azure solutions, Python, SQL, Bash, HDP, Hadoop, Spark, Hive, Pig, Sqoop, PySpark, Scala, PostgreSQL, Cassandra, ARM templates, XML, JSON, Docker, Kubernetes, Bitbucket, Azure DevOps, ELK, JIRA, Agile, Looker.

**Guardian Life Insurance limited, Bethlehem, PA Feb 2021– Dec 2021**

**AWS Big Data Engineer**

**Responsibilities:**

* Design and develop data pipelines to decrypt and encrypt [individual and zipped] files in **s3 bucket** using **boto3**. Transformed data using **PySpark** transformations.
* Developed Airflow Dag’s with **Python** for Task Schedule and automated daily run.
* Developed multiple **Adhoc SQL Query** and complicated views as per business requirements in **Snowflake** using complicated functions like split\_by\_table, map\_agg, and array agg reducing the query latency by 82%.
* Designed and developed internal data infrastructure and analytics support from the ground up by deploying custom **Python** Batch **ETL** subsystems using **Docker**, establishing a data warehouse on **AWS redshift.**
* Migrated on-premises **SAP ASE** database to **AWS EC2** using **AWS** database migration service (DMS).
* Created **Python** scripts for data access and analysis to aid in the process and system monitoring, and reporting.
* Developed **ETL** pipelines using **SSIS** to acquire call volume data from multiple sources and build a dataset to be used for call volume forecast by data scientists.
* Integrated **SAP SRM** with **SAP IBO** for improved process monitoring for a major telecom company with 100% defect-free code at the time of Go-Live.
* Implemented **SAP CRM** (Customer Relationship Management) Service Module.
* Developed **ERMS** **(Employee Management Response System)** which helped to reduce downtime on tickets handled by the customer service agent.
* Engineered and executed end-to-end ETL pipelines using SSIS to provide a platform for the alternative dataset.
* Designed dashboard using **Tableau** to monitor the movement of customers between deciles based on their propensity score to inform businesses to take necessary actions.
* Implemented **POC** to develop a Data **OPS** solution for AWS Services using S3, Glue, Lambda, **Code Pipeline**, and **Terraform.**
* Optimized the Hive tables using optimization techniques like **partitions** and **bucketing** to provide better performance with **HiveQL** queries.
* Developed intent classification model using **python NLP** libraries **spacy**, **NLTK**, and **Genism** to identify the customer intent for smart routing of email to service reps based on tags like inquiry, feedback, and dissatisfaction.

**Environment**: Linux, AWS EC2, S3, Python 3.6, R, pandas, R, Hadoop 2.7.4, SAP CRM, SRM, NLP, spacy, NLTK, Genism, SQL, Tableau.

**Santander Bank, NYC, NY Nov 2019 – Jan 2021**

**Data Engineer**

**Responsibilities:**

* Involved in developing roadmap for migration of enterprise data from multiple data sources like **SQL Server**, provider databases into **S3** which serves as a centralized datahub across the organization.
* Loaded and transformed large sets of **structured** and **semi structured** data from various downstream systems.
* Developed **ETL** pipelines using **Spark** and **Hive** for performing various business specific transformations.
* Building Applications and automating the pipelines in **Spark** for bulk loads as well as Incremental Loads of various Datasets.
* Worked closely with our team’s **data scientists** and consumers to shape the datasets as per the requirements.
* Automated the data pipeline to **ETL** all the Datasets along with full loads and **incremental loads** of data.
* Utilized **AWS** services like **EMR, S3, Glue** **Metastore** and **Athena** extensively for building the data applications.
* Worked on building input adapters for data dumps from **FTP** Servers using **Apache spark**.
* Wrote **spark** **applications** to perform operations like data inspection, cleaning, load and transforms the large sets of structured and semi-structured data.
* Developed **Spark** with **Scala** and **Spark-SQL** for testing and processing of data.
* Reporting the spark job stats, Monitoring and Running Data Quality Checks are made available for each Datasets.
* Used SQL Programming Skills to work around the Relational SQL Databases

**Environment:** AWS Cloud Services, Apache Spark, Spark-SQL, Unix, Kafka, Scala, SQL Server.

**Extarc Software Solutions Pvt. Ltd, India May 2017 – October 2019**

**Hadoop Developer**

**Responsibilities:**

* Developed **ETL jobs** using Spark -Scala to migrate data from Oracle to new **MySQL** tables.
* Rigorously exerted **Spark -Scala (RDD’s, Data frames, Spark SQL)** and **Spark** - **Cassandra** -Connector APIs for various tasks (Data migration, Business report generation etc.)
* Developed **Spark Streaming** application for real time sales analytics.
* Prepared an **ETL framework** with the help of **Sqoop, pig** and **hive** to be able to frequently bring in data from the source and make it available for consumption.
* Processed **HDFS** data and created external tables using Hive and developed scripts to ingest and repair tables that can be reused across the project.
* Engineered complex data pipelines using tools such as **Databricks,** processing terabytes of data to drive decision-making.
* Analyzed the source data and handled efficiently by modifying the data types. Worked on **excel sheet, flat files, CSV files** to generated **PowerBI** ad-hoc reports.
* Analyzed the SQL scripts and designed the solution to implement using **PySpark**.
* Extracted the data from other data sources into **HDFS** using **Sqoop**.
* Handled importing of data from various data sources, performed transformations using **Hive, MapReduce,** loaded data into **HDFS**.
* Extracted the data from **MySQL** into **HDFS** using **Sqoop**.
* Implemented automation for deployments by using **YAML** scripts for massive builds and releases.
* **Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka and Sqoop.**
* Implemented Data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, Partitioning, distributed cache to improve the performance of MapReduce jobs.
* Worked on **GIT** to maintain source code in Git and GitHub repositories.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Cassandra, YAML, ETL.

**Triniti Advanced Software Labs, Hyderabad, India May 2014 – April 2017**

**Data Warehouse Consultant**

**Responsibilities:**

* Worked with Stakeholders regarding business requirements, functional specifications, and enhancements, based on the business needs created technical design and functional specification documents.
* Active participation in **weekly calls** with **data modeling** and analyst teams to understand and work on any new requirements.
* Analyzing data from source systems to design the solution for the business requirement.
* Developed Complex Mappings in Informatica using Power Center transformations (**Source Qualifier, Joiner, Lookups, Filter, Router, Aggregator, Expression, XML** **Update and Sequence generator transformations**), Mapping Parameters/Variables, Parameter files, SQL overrides, Transformation Language.
* Data profiling the source files and developing data model and mappings for smaller requirements.
* Implemented **CDC, SCD2, SCD1** Delta load, Snapshot and transactional fact tables, headers and footers to Flat File, File list.
* Developed Unix scripts for **SFTP** file transfers and Target table truncate operations.
* Implemented partitioning at database level for better performance.
* Scheduled the Workflows to run on a daily and weekly basis using **Control-M** Scheduling tool.
* Provided support to the **QA team** for various testing phases of ETL development.
* Involved in unit testing and Unit test plan document preparation.
* Documenting mapping specifications, **STM,** Unit Test cases, procedure, results.
* Implemented **Push Down Optimization (PDO)** for better performance when source data is huge.

**Environment**: Informatica PowerCenter 9.5/9.1, Oracle 11g, Flat Files, COBOL Files, Erwin, Control-M, SQL, PL/SQL, Shell Scripting.