**Professional Summary**

* Data engineering professional having 9+ years of experience in a variety of data platforms, with hands on experience in Big Data Engineering and Data Analytics.
* Strong Expertise with tools in **Hadoop** Ecosystem and Big data stack which includes **HDFS**, **Spark**, **Airflow**, **Pyspark**, **MapReduce**, **Kafka**, **Hive**, **Oozie**, **Zookeeper**, **Ambari**, **HBase**, and **Impala**.
* Worked on analyzing Hadoop cluster and different big data analytic tools including Pig HBase database and Sqoop.
* Experience in layers of **Hadoop** Framework - Storage (**HDFS**), Analysis (**Pig** **and** **Hive**), Engineering (Jobs and Workflows), extending the functionality by writing custom **UDFs**.
* Hands-on experience with **Hadoop** architecture and various components such as **Hadoop** File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and **Hadoop** **MapReduce** programming.
* Having Proficient experience in various big data technologies like **Hadoop**, **Apache** **Nifi**, **Hive** **Query** Language, **HBase** **NoSQL** database, **Sqoop**, **Spark**, **Scala**, **OOZIE** and **Pig** etc.), Oracle Database and **Unix** shell Scripting technologies.
* Design and implement database solutions in **Azure** **SQL** **Data** **Warehouse**, **Azure** **SQL**
* Reviewed the **HDFS** usage and system design for future scalability and fault-tolerance. Installed and configured **Hadoop** **HDFS**, **MapReduce**, **Pig**, **Hive**, **Sqoop**.
* Strong experience in **Teradata**, **Informatica** , **Python**, **UNIX** shell scripting for processing large volumes of data from varied sources and loading into databases like **Teradata**, Oracle.
* 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**, **NoSQL** **DB**).
* Created **Azure** **SQL** database, performed monitoring and restoring of **Azure** **SQL** database. **Performed** **migration** of **Microsoft** **SQL** server to **Azure** **SQL** database.
* Extensively used **Python** Libraries **PySpark**, **Pytest**, **Pymongo**, **cxOracle**, **PyExcel**, **Boto3**, **embedPy**, **Pandas**, **NumPy**.
* Expert in working with **Hive** data warehouse tool-creating tables, data distribution by implementing partitioning and bucketing, writing, and optimizing the **HiveQL** queries.
* Extensive Shell/**Python** scripting experience and setup production systems on **UNIX**/**LINUX** environment.
* Experience in providing application support for **Jenkins**.
* Primarily involved in Data Migration using **SQL**, **SQL** **Azure**, **Azure** **Storage**, and **Azure** **Data** **Factory**, **SSIS**, **PowerShell**.
* Experience in working with Amazon Web Services (**AWS**) Using **EC2** for computing, **Redshift** Spectrum, **Glue** and **S3** as storage mechanism.
* Strong working knowledge of Amazon Web Services (**AWS**) Cloud Platform which includes services like **EC2**, **S3,** **VPC**, **ELB**, **IAM**, **DynamoDB**, **Cloud** **Front**, **Cloud** **Watch**, **Route** **53**, Elastic Beanstalk (**EBS**), **Auto** **Scaling**, Security Groups, **EC2** Container Service (**ECS**), Code Commit, Code Pipeline, Code Build, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Red **shift**, **CloudWatch**, **Cloud** **Formation**, **Cloud** **Trail**, **Ops** **Works**, **Kinesis**, **IAM**, **SQS,** **SNS**, **SES.**
* Worked in agile projects delivering end to end continuous integration/continuous delivery pipeline by Integration of tools like **Jenkins** and **AWS** for VM provisioning.
* Good Knowledge on Cloudera distributions and in Amazon simple storage service (**Amazon** **S3**), **AWS** **Redshift**, **Lambda** and Amazon **EC2**, Amazon **EMR**.
* Excellent understanding of **Hadoop** Architecture and good Exposure in **Hadoop** components like **Hadoop** **Map** **Reduce**, **HDFS**, **HBase**, **Hive**, **Sqoop**, **Cassandra**, **Kafka** and **Amazon** **Web** **services** (**AWS)** **API** **test**, document and monitor by Postman which is easily integrate the tests into your build automation.
* Worked on Amazon Web service (**AWS**) to integrate **EMR** wif Spark 2 and S3 storage and Snowflake.
* Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (**AWS**) on **EC2**.
* Handled **AWS** Management Tools as Cloud watch and Cloud Trail. Stored tech log files in **AWS** S3. Used versioning in **S3** buckets where the highly sensitive information is stored.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop Eco System** | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala,  Apache Flume, Apache Storm, Apache Airflow, HBase |
| **Programming Languages** | PL/SQL, SQL, Python, Scala, PySpark, C, C++ |
| **Cluster Mgmt& monitoring** | CDH 4, CDH 5, Horton Works Ambari 2.5 |
| **Data Bases** | MySQL, SQL Server, Oracle, MS Access |
| **NoSQL Data Bases** | MongoDB, Cassandra, HBase |
| **Workflow mgmt. tools** | Oozie, Apache Airflow |
| **Visualization & ETL tools** | BananaUI, D3.js, Informatica, Talend |
| **Cloud Technologies** | Azure, AWS |
| **IDE’s** | Eclipse, Jupyter notebook, Spyder, PyCharm, IntelliJ |
| **Version Control Systems** | Git, SVN |
| **Operating Systems** | Unix, Linux, Windows |

**Professional Experience**

**Client: Lending Tree, Charlotte, NC Oct2020- Present**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Performed data analysis and developed analytic solutions. Data investigation to discover correlations / trends and the ability to explain them.
* Used **Spark** **SQL** functions to move data from stage **hive** tables to fact and dimension tables.
* Using **HBase** to store majority of data which needs to be divided based on region.
* Used **Oozie** and **Zookeeper** operational services for coordinating cluster and scheduling workflows.
* Used **Zookeeper**to provide coordination services to the **cluster**. Experienced in managing and reviewing **Hadoop** log files.
* Involved in creating **Unixshell** scripts for **Informatica** workflow execution.
* Created data bricks notebooks using **Python**(PySpark),Scala and Spark SQL for transforming the data that is stored in **Azure** Data Lake stored Gen2 from Raw to Stage and Curated zones.
* Built numerous technology demonstrators using Confidential Edison Arduino shield using **Azure** EventHub and Stream Analytics, integrated with PowerBI and **Azure** ML to demonstrate the capabilities of **Azure** Stream Analytics.
* Developed JSON Scripts for deploying the Pipeline in **Azure** Data Factory (ADF) that process the data using the **Sql** Activity.
* Responsible for writing Autosys **JILs** and scheduling the **Informatica** workflows on the **Autosys** server
* Developed **JSON** Scripts for deploying the Pipeline in **Azure** **Data** **Factory** (**ADF**) that process the data using the Sql Activity.
* Build Infrastructure as a code using terraform and cloud formation templates (**JSON** &**YAML**).
* Experienced in troubleshooting errors in **Hbase** Shell/API, **Pig**, **Hive** and **map Reduce**.
* Experience in using **Zookeeper** and **Oozie** operational services to coordinate clusters and scheduling workflows. Migrated SQL Server and Oracle database to Microsoft **Azure** Cloud.
* Developed **stored procedures/views in Snowflake** and use in **Talend** for loading Dimensions and Facts.
* Experienced in configuring and administering the **Hadoop** Cluster using major **Hadoop** Distributions like Apache **Hadoop** and **Cloudera**.
* Involved in importing real time Jet.com data to **HDFS** from **Kafka** and implemented and scheduled hourly runs using **automic**/**uc4**.
* Involved in Importing and exporting data from **HDFS** using **Sqoop**, resolution of access issues, performance issues and Patch/upgrade related issues.
* Develop batch processing solutions by using **Data** **Factory** and **Azure** **Data** **bricks**
* Developed merge scripts to **UPSERT** data into **Snowflake** from an ETL source.
* Worked on **Docker containers** snapshots, attaching to a running container, removing images, managing Directory structures and managing containers.

**Environment:** Hadoop, Map Reduce, Azure, HDFS, Hive,Snowflake, Spring Boot, Swamp, Data Bricks, Data Lake, Sqoop, Oozie, SQL, Kafka, Spark, Scala, GitHub, Informatica ,Python, Talend Big Data Integration, Solr, Impala.

**Client: Centene Corporation, St Louis, Missouri Sept2019- Sept2020**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Used Data Frame API in Scala for converting the distributed collection of data organized into named columns, developing predictive analytic using Apache **Spark** **Scala** **APIs**.
* Involved in designing and deploying multi-tier applications using all the **AWS** services like (**EC2**, **Route53**, **S3**, **RDS**, Dynamo DB**,** **SNS**, **SQS**, **IAM**) focusing on high-availability, fault tolerance, and auto-scaling in **AWS** Cloud Formation.
* Supporting Continuous storage in **AWS** using Elastic Block Storage, S3. Created Volumes and configured Snapshots for **EC2** instances.
* Developed and deployed stacks using **AWS** Cloud Formation Templates (**CFT**) and **AWS** Terraform.
* Implement the Kafka to hive streaming process flow and batch loading of data into MariaDB using Apache Nifi. Implementations of generalized solution model using **AWS** SageMaker.
* Create develop and test environments of different applications by provisioning Kubernetes clusters on **AWS** using **Docker**, **Ansible**, and **Terraform**.
* Worked with **informatica** support for fixing **Informatica** Linux server issues.
* Worked on **ETL** pipeline to source these tables and to deliver this calculated ratio data from **AWS** to Datamart (**SQL** **Server**) & Credit Edge server.
* Experience in using and tuning relational databases (e.g. **Microsoft** SQL **Server**, **Oracle**, **MySQL**) and columnar databases (e.g. **Amazon** **Redshift**, **Microsoft** **SQL** **Data** **Warehouse**).
* Handled **AWS** Management Tools as Cloud watch and Cloud Trail.
* Stored the log files in **AWS** **S3**. Used versioning in S3 buckets where the highly sensitive information is stored. Integrated **AWS** Dynamo **DB** using **AWS** lambda to store the values of items and backup the **DynamoDB** streams
* Automated Regular **AWS** tasks like snapshots creation using **Python** scripts. Install and configure Apache Airflow for **AWS** S3 bucket and created **dags** to run the Airflow
* Prepared scripts to automate the ingestion process using **Pyspark** and **Scala** as needed through various sources such as API, **AWS** **S3**, **Teradata** and **Redshift**.
* Worked on **AWS** Data pipeline to configure data loads from **S3** to into **Redshift**.
* Worked on **ETL** Migration services by developing and deploying **AWS** Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.
* Extensive expertise using the core **Spark** **APIs** and processing data on an **EMR** cluster.
* Developed **Python** scripts using both Data frames/**SQL**/Data sets and **RDD/MapReduce** in **Spark** for Data Aggregation, queries and writing data back into **OLTP** system through **Sqoop**.
* Developed **Hive** queries to pre-process the data required for running the business process.
* Created **HBase** tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, **NoSQL** and a variety of portfolios.

**Environment**: Hortonworks, Hadoop, HDFS, AWS Glue, Pig, Sqoop, Hive, NoSQL, Terraform, HBase, Shell Scripting, Python, Scala, Spark, SparkSQL, SQL Server, Informatica ,ETL.

**Client: CUMMINS, Columbus, Indiana Sept2017-August2019**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Responsible for building scalable distributed data solutions using **Hadoop**.
* Used spark steaming **APIs** to perform necessary transformations for building the common learner data model which gets data from **Kafka** in near real time and persists into Hive.
* Developed **Spark** scripts by using **Python** as per the requirements.
* Developed real time data pipeline using **Spark** to ingest customer events/activity data into Hive and **Cassandra** from **Kafka**.
* Design and implement database solutions in **Azure** SQL Data Warehouse, **Azure** SQL.
* 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).
* Extensively involved in performance tuning of the **Informatica** **ETL** mappings by using the caches and overriding the **SQL** queries and by using Parameter files
* Design & implement migration strategies for traditional systems on **Azure** (Lift and shift/**Azure** Migrate, other third-party tools.
* Used various sources to pull data into **Power** **BI** such as **SQL** Server, Excel, Oracle, **SQL** **Azure** etc.
* Propose architectures considering cost/spend in **Azure** and develop recommendations to right-size data infrastructure.
* Reviewed **Informatica** mappings and test cases before delivering to Client.
* Design Setup maintain Administrator the **Azure** **SQL** Database, **Azure** Analysis Service, **Azure** SQL Data warehouse, **Azure** Data Factory, **Azure** **SQL** Data warehouse.
* Worked on reading and writing multiple data formats like **JSON,** **AVRO**, **Parquet**, **ORC** on **HDFS** using Pyspark.
* Designed, developed, and did maintenance of data integration in **Hadoop** and **RDBMS** environment with both traditional and non-traditional source system as well as **RDBMS** and **NoSQL** data stores for data access and analysis.
* Involved in recovery of **Hadoop** clusters and worked on cluster size of 310 nodes.
* Worked on creating **Hive** tables, loading, and analyzing data using Hive queries.

**Environment**: Python, Bigdata, Hadoop, HBase, Hive, Azure, Cassandra, Spark, Pyspark, Cloudera, Kafka, Sqoop, Unix Shell scripting, GitHub, SQL.Informatica, Power BI.

**Client: Ditech Home Loans, Tempe, AZ Aug2015 - Aug2017**

**Role: Data Engineer**

**Responsibilities:**

* Built S3 buckets and managed policies for S3 buckets and used **S3** glacier for storage and backup on **AWS**
* Designed, built, and coordinate an automated build & release **CI**/**CD** process using GitLab, Jenkins and Puppet on hybrid IT infrastructure.
* Involved in designing and developing **Amazon** **EC2**, **Amazon** **S3**, **Amazon** **RDS**, **Amazon** **Elastic** Load Balancing, **Amazon** **SWF**, **Amazon** **SQS**, and other services of the **AWS** infrastructure.
* Running build jobs and integration tests on Jenkins Master/Slave configuration.
* Conduct systems design, feasibility and cost studies and recommend cost-effective cloud solutions such as Amazon Web Services (**AWS**).
* Used **dynamic** cache memory and index cache to improve the performance of **Informatica** server
* Involved in maintaining the reliability, availability, and performance of **Amazon** Elastic Compute Cloud (**Amazon** **EC2**) instances.
* Managed Servers on the Amazon Web Services (**AWS**) platform instances using Puppet configuration management.
* Integrated services like **GitHub**, **AWS** Code **pipeline**, **Jenkins**, and **AWS** Elastic Beanstalk to create a deployment pipeline.
* Involved in complete **SDLC** life cycle - **Designing**, **Coding**, **Testing**, **Debugging** and Production Support.
* Coordinate/assist developers with establishing and applying appropriate branching, labeling/naming conventions using **Git**.
* Analyzed Bulk Load option, Third party Loaders suggested by **Informatica**.
* Used Kubernetes to deploy scale, load balance, scale and manage docker containers.
* Worked on **JIRA** for defect/issues logging & tracking and documented all my work using CONFLUENCE
* Branching, Merging, Release Activities on Version Control Tool **GIT**. Used **GitHub** as version control to store source code and implemented Git for branching and merging operations.

**Environment**: Jenkins, JIRA, Maven, GIT, AWS EMR/EC2/S3/Redshift ,Informatica **,** Cassandra, Web logic, Unix Shell Scripting, SQL, Kubernetes, docker.

**Client: Ascena retail group, India May2013 - May2015**

**Role: Data Analyst**

**Responsibilities:**

* Collaborated with multiple business units to gather requirements and develop data reporting infrastructure to provide real-time business insights.
* Modeled a new data mart on **SQL** server to hold subset of data from multiple data sources by diverse use of SQL queries with working knowledge of **RDBMS** and **ETL** tools.
* Built the infrastructure required for optimal extraction, transformation and loading of data from a wide variety of data sources using **SQL**.
* **Azure** Data Factory (**ADF**), Integration Run Time (**IR**), File System Data Ingestion, Relational Data Ingestion.
* Worked in mixed role **DevOps** **Azure** Architect/System Engineering, network operations and data engineering.
* Used dynamic cache memory and index cache to improve the performance of **Informatica** server
* Strong experience with **Informatica** Designer, Workflow Manager, Workflow Monitor, Repository Manager.
* Development and maintenance of data pipeline on **Azure** Analytics platform using **Azure** Databricks, **PySpark, Python, Pandas** and **NumPy** libraries.
* Working with data ingestions from multiple sources into the **Azure** SQL data warehouse
* Transforming and loading data into **Azure** SQL Database
* Developing purging scripts and routines to purge data on **Azure** SQL Server and **Azure** Blob storage.
* Maintaining data storage in **Azure** Data Lake.
* Created **Azure SQL** database, performed monitoring and restoring of **Azure** **SQL** database. Performed migration of Microsoft **SQL** server to **Azure SQL** database.
* Extract Transform and Load data from Sources Systems to **Azure** Data Storage services using a combination of **Azure** **Data** **Factory**, **T**-**SQL**, **Spark** **SQL** and **U-SQL** **Azure** **Data** **Lake** **Analytics**. Data Ingestion to one or more **Azure** **Services** - (**Azure** **Data** **Lake**, **Azure** **Storage**, **Azure** **SQL**, **Azure** **DW**) and processing the data in **Azure** **Databricks**.

**Environment**: Azure, Informatica , Oracle server, PL/SQL, Linux, Python.

**EDUCATION: Bachelor’s in CSE JNTUH 2009-2013**