**Anusha k**

**(980)222-0603 |** [**anushak9631@gmail.com**](mailto:anushak9631@gmail.com)

**SUMMARY:**

* Having 10 years of IT experience in a variety of industries working on Big Data technology using technologies such as Cloudera and Hortonworks distributions. **Hadoop** working environment includes **Hadoop**, **Spark**, **Map Reduce**, **Kafka**, **Pig**, **Hive**, **Ambari**, **Sqoop**, **HBase** and **Impala**.
* 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**.
* Hands-on experience in **data structure**, design and analysis using **Machine Learning Technics and modules in PYTHON, R.**
* Good experience working with various data analytics and big data services in **AWS Cloud** like **EMR**, **Redshift**, **S3**, **Athena**, **Glue** etc.,
* 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.
* Experience in architecting, designing, and building **distributed** **data pipelines.**
* Deep knowledge of **troubleshooting** and **tuning Spark** applications and **Hive** **scripts** to achieve optimal performance.
* Works on loading data into **Snowflake DB** in the cloud from various sources.
* Experience on Migrating SQL database to **Azure Data Lake**, **Azure data lake Analytics**, **Azure SQL Database, Data Bricks** and **Azure SQL Data warehouse** and Controlling and granting database access and Migrating On premise databases to **Azure Data lake store** using **Azure Data factory.**
* Design and develop **hive**, **HBase** data structure and **Oozie** **workflow**.
* Experience developing **Kafka** producers and **Kafka** Consumers for streaming millions of events per second on streaming data.
* Expertise with **Python**, **Scala** and **Java** in Design, Development, Administrating and Supporting of large-scale distributed systems.
* Experience in **GCP Dataproc, GCS, Cloud functions, BigQuery.**
* Experience working with **NoSQL** **databases** like **Cassandra** and **HBase** and developed **real-time** read/write access to very large datasets via **HBase**.
* Extensive usage of **Azure Portal**, **Azure PowerShell**, Storage Accounts, Certificates and **Azure Data** Management.
* Significant experience writing custom **UDF’s** in **Hive** and custom Input Formats in **Map Reduce**.
* Schedule nightly batch jobs using **Oozie** to perform schema validation and IVP transformation at larger scale to take the advantage of the power of **Hadoop**.
* Imported the customer data into **Python using Pandas** libraries and performed various data analysis - found patterns in data which helped in key decisions.
* Good experience is designing and implementing end to end data security and governance within **Hadoop** Platform using **Kerberos**.
* Worked with **real-time** data processing and **streaming** techniques using **Spark streaming** and **Kafka**.
* Experience in moving data into and out of the **HDFS** and **Relational Database Systems (RDBMS)** using **Apache** **Sqoop**.
* Experience in Designing, Architecting and implementing **scalable** cloud-based web applications using **AWS**.
* Having good knowledge in writing **Map Reduce** jobs through **Pig**, **Hive**, and **Sqoop**.
* Strong experience **productionalizing** end to end data pipelines on **Hadoop** platform.
* Experience using various **Hadoop** Distributions **(Cloudera, Hortonworks, Amazon AWS EMR)** to fully implement and utilize various **Hadoop** services.
* Configure **Zookeeper** to coordinate and support **Kafka**, **Spark**, **Spark** Streaming, **HBase** and **HDFS**
* Used **Oozie** and **Zookeeper** operational services for coordinating cluster and scheduling workflows.
* Experience working with **NoSQL** databases like **MongoDB**, **Cassandra** and **HBase**.
* Used **Hive** extensively for performing various data analytics required by business teams.
* Solid experience in working various data formats like **Parquet**, **Orc**, **Avro**, **Json** etc.,

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Data/Hadoop Technologies** | Map Reduce, Spark, SparkSQL, Spark Streaming, Kafka,  PySpark, Pig, Hive, HBase, Flume, Yarn, Oozie, Zookeeper, Hue,  Ambari Server |
| **Languages** | HTML5,DHTML, WSDL, CSS3 ,C, C++, XML,R/R Studio, SAS  Enterprise Guide, SAS ,R (Caret, Weka, ggplot) , Perl, MATLAB,  Mathematica, FORTRAN, DTD, Schemas, Json, Ajax, Java, Scala,  Python (NumPy, SciPy, Pandas, Gensim, Keras), Java Script, Shell  Scripting |
| **SQL Databases** | Cassandra, HBase, MongoDB |
| **Design Tools** | HTML, CSS, JavaScript, JSP, jQuery, XML |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Azure Data bricks, Eclipse, NetBeans. |
| **Cloud** | AWS, MS Azure, GCP, Snowflake |
| **Development**  **MeMethodologies** | 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), Crystal reports  XI, SSRS, cognos. |
| **Databases** | Microsoft SQL Server, MySQL , Oracle, DB2, Teradata, Netezza |
| **Operating Systems** | All versions of Windows, UNIX, LINUX, Macintosh HD, Sun Solaris |

**WORK EXPERIENCE:**

**Client: Mayo Clinic, Rochester, MN May 2022 – till now**

**Role: Senior Big Data Engineer**

**Responsibilities:**

* Developed **PIG** Latin scripts for the analysis of semi structured data.
* Developed cloud strategies using **GCP (for its PAAS)**
* Got involved in migrating on prem Hadoop system to using **GCP (Google Cloud Platform).**
* 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.
* Design & implement **Spark** **SQL** tables, **Hive** scripts job with stone branch for scheduling and create work flow and task flow.
* Build data pipelines in **airflow** in **GCP** for ETL related jobs using different airflow operators.
* Excellent understanding of **Hadoop** architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and **Map Reduce** programming paradigm.
* Leveraged **Google Analytics** to track website performance and user behavior, resulting in a 20% increase in website conversion rates.
* Conducted in-depth **Adobe Analytics** (Omniture) analyses to identify customer journey bottlenecks and recommended UX improvements, leading to a 15% reduction in bounce rates.
* Utilized **SQL** to retrieve and analyze large datasets, creating custom reports for marketing campaigns that improved ROI by 10%.
* Collaborated with cross-functional teams to integrate data from various sources, enhancing data-driven decision-making processes.
* Automated weekly and monthly reports using SQL scripts, saving an average of 10 hours per week in manual data processing.
* This example demonstrates how you can showcase your skills and provide concrete examples of how you've applied them in your previous roles. Tailor your resume to match the specific job you're applying for, emphasizing the skills and experiences most relevant to that position.
* Bulk loading and unloading data into **Snowflake** tables using COPY command.
* Working in big data technologies like **spark**, **Scala**, **Hive**, **Hadoop** cluster (Cloudera platform).
* Making a data pipelining with help Data Fabric job, **SQOOP**, **SPARK**, **Scala** and **KAFKA**. Parallel working in data side oracle and MYSQL server for data designing to source to target.
* Expertise in data transformation & analysis using **SPARK,** **PIG, HIVE**
* We generally used partitions and bucketing for data in **hive** to get query faster. This part of **hive** optimization
* Developed **Kafka** consumer API in **Scala** for consuming data from **Kafka** topics.
* Responsible to manage data coming from different sources through **Kafka**.
* Good Exposure on **Map Reduce** programming using Java, **PIG** Latin Scripting and Distributed Application and HDFS.
* Validate the data feed from the source systems to **Snowflake DW** cloud platform.
* Used cloud shell SDK in **GCP** to configure the services Data Proc, Storage, BigQuery
* Involved in Importing and exporting data from HDFS using **Sqoop**, resolution of access issues, performance issues and Patch/upgrade related issues.
* Experience in working with **Map Reduce** programs using Apache **Hadoop** for working with Big Data
* Involved in creating **Hive**QL on **HBase** tables and importing efficient work order data into **Hive** tables
* Extending **Hive** and **Pig** core functionality by writing custom **UDFs, UDTF and UDAFs.**
* Used **Oozie** and **Zookeeper** operational services for coordinating cluster and scheduling workflows.
* Write programs using **Spark** to move data from Storage input location to output location by running data loading, validation, and transformation to the data
* Integrated and automated data workloads to **Snowflake Warehouse.**
* Ensure ETL/ELT’s succeeded and loaded data successfully in **Snowflake DB.**
* Migrated **Map reduce** jobs to **Spark** jobs to achieve better performance.
* Handled importing of data from various data sources, performed transformations using **Hive**, **MapReduce**, loaded data into HDFS and extracted data from MYSQL into HDFS vice-versa using **Sqoop**.
* Used **scala** function, dictionary and data structure (array, list, map) for better code reusability
* Based on Development, we need to do the Unit Testing.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on **GCP**.
* Developed Java **Map Reduce** programs for the analysis of sample log file stored in cluster.
* Working in relational SQL and NoSQL databases, including Oracle, **Hive**, **Sqoop** and **Hbase**

**Environment:** HDFS, Hive, Spark, Snowflake, GCP, BigQuery, Data Proc, Data Storage, Airflow, Linux, Kafka, python, Stone branch, Cloudera, Oracle 12c, PL/SQL, Unix, Json and Parquet File systems.

**Client: Nationwide, Columbus, OH Jan 2021 – April2022**

**Role: Senior Data Engineer**

**Responsibilities:**

* Utilized **Agile and Scrum** methodology for team and project management.
* Selected and generated data into csv files and stored them into **AWS S3** by using **AWS EC2** and then structured and stored in **AWS Redshift**.
* Experience in using **Kafka** and **Kafka** brokers to initiate **spark** context and processing **livestreaming**.
* Developed **Scala** scripts, UDFs using both Data frames/**SQL**/Data sets and RDD in **Spark** for Data Aggregation, queries and writing data back into OLTP system through **Sqoop**.
* Involved in **HBASE** setup and storing data into **HBASE**, which will be used for further analysis.
* Closely worked with **Kafka** Admin team to set up **Kafka** cluster and implemented **Kafka** producer and consumer applications on **Kafka** cluster setup with help of **Zookeeper.**
* Experience in using **Zookeeper** and **Oozie** operational services to coordinate clusters and scheduling workflows
* Good Exposure on **Map Reduce** programming using Java, **PIG** Latin Scripting and Distributed Application and **HDFS**.
* Migrated users from Teradata to snowflake and created the roles and assigned the users to the roles.
* Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
* Partnered with ETL developers to ensure that data is well cleaned and the data warehouse is up-to-date for reporting purpose by **Pig**.
* Developed **stored procedures/views in Snowflake** and use in **Talend** for loading Dimensions and Facts.
* Developed custom **Kafka** producer and consumer for different publishing and subscribing to **Kafka** topics.
* Wrote **Scala** scripts to make **spark** streaming work with **Kafka** as part of **spark** **Kafka** integration efforts.
* Performed data preprocessing and feature engineering for further predictive analytics using **Python Pandas**.
* Used **Spark** Streaming to receive real time data from the **Kafka** and store the stream data to **HDFS** using **Python** and **NoSQL** databases such as **HBase** and **Cassandra**
* Generated report on predictive analytics using **Python** and **Tableau** including visualizing model performance and prediction results.
* Hands on experience in Capturing data from existing relational databases (**Oracle, MySQL, SQL** and **Teradata**) that provide **SQL** interfaces using **Sqoop**.
* Used **PySpark** and **Pandas** to calculate the moving average and RSI score of the stocks and generated them into data warehouse.
* Installed and configured **Hadoop** **Map Reduce**, **HDFS**, developed multiple **Map Reduce** jobs in java and **Scala** for data cleaning and preprocessing.
* Used **HBase**/Phoenix to support front end applications that retrieve data using row keys
* Designed and executed **Oozie** workflows in a manner that allowed for scheduling **Sqoop** and **Hive** job actions to extract, transform and load data
* Exploring with **Spark** to improve the performance and optimization of the existing algorithms in **Hadoop** using **Spark context, Spark-SQL,postgreSQL,Data Frame,OpenShift, Talend,pair RDD's**
* Responsible for importing data from **Postgres** to **HDFS, HIVE using SQOOP tool**.
* Experienced in migrating **Hive QL** into Impala to minimize query response time.
* Involved in Functional Testing, Integration testing, Regression Testing, Smoke testing and performance Testing. Tested **Hadoop**, **Map Reduce** developed in **python**, **pig**, **Hive**.
* Created functions and assigned roles in **AWS Lambda** to run **python** scripts, and **AWS Lambda** using java to perform event driven processing. Created Lambda jobs and configured Roles using **AWS CLI.**
* Good working experience on **Spark** (**spark streaming, spark SQL**) with **Scala** and **Kafka**. Worked on reading multiple data formats on **HDFS** using **Scala**.
* Expertise in using **Docke**r to run and deploy the applications in multiple containers like **Docker Swarm** and **Docker Wave.**
* Developed complex **Talend ETL jobs** to migrate the data from **flat files** to database. Pulled files from **mainframe into Talend** execution server using multiple **ftp** components.
* Used **Sqoop** to channel data from different sources of **HDFS** and **RDBMS**
* Architect and design server less application **CI/CD** by using **AWS Server less (Lamda**) application model.

**Environment**: Hdfs, Hive, Spark, Kafka, linux, Python, Numpy, Pandas, Tableau, GitHub, AWS EMR/EC2/S3/Redshift, Lambda, Pig, Map Reduce, Cassandra, Snowflake, Unix, Shell Scripting, Git.

**Client: AgFirst, Columbia, SC Dec 2018 – Dec 2020**

**Role:  Big Data Engineer**

**Responsibilities:**

* Used **Spark** streaming to receive real time data from the **Kafka** and store the stream data to **HDFS** using **Scala** and No**Sql** databases such as **HBase** and Cassandra.
* Configured **Oozie** workflow to run multiple **Hive** and **Pig** jobs which run independently with time and data availability.
* Implemented Apache Drill on **Hadoop** to join data from **SQL** and No **SQL** databases and store it i Configured **Spark** streaming to receive real time data from **Kafka** and store the stream data to **HDFS** using **Scala**.
* Effectively used **Sqoop** to transfer data from databases (**SQL, Oracle**) to **HDFS, Hive**.
* Used **Hive** to analyse data ingested into **HBase** by using **Hive**-**HBase** integration and compute various metrics for reporting on the dashboard.
* Transforming business problems into Big Data solutions and define Big Data strategy and Roadmap. Installing, configuring, and maintaining Data Pipelines
* Developed **Sqoop** scripts to import and export data from relational sources by handling incremental data loading on the customer transaction data by date.
* Responsible for running **Hadoop** streaming jobs to process terabytes of xml's data, utilized cluster co-ordination services through **Zookeeper**.
* Wrote **AZURE POWERSHELL** scripts to copy or move data from local file system to HDFS Blob storage.
* Used **Oozie** workflow engine to manage independent **Hadoop** jobs and to automate several types of **Hadoop** such as java **Map Reduce**, **Hive** and **Sqoop** as well as system specific jobs
* Creating Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure **SQL**, Blob storage, Azure **SQL** Data warehouse, write-back tool and backwards.
* Experienced of building **Data Warehouse** in **Azure platform** using **Azure data bricks** and **data** **factory**.
* Experience with leveraging **Hadoop** ecosystem components including **Pig** and **Hive** for data analysis, **Sqoop** for data migration, **Oozie** for scheduling and **HBase** as a No**SQL** data store.
* Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services. Knowledge of U**SQL**
* Troubleshooting the **Azure Development**, configuration and Performance issues.
* Worked on results from **Kafka** server output successfully.
* Utilized **Spark**, **Scala**, **Hadoop**, **HBase**, Cassandra, MongoDB, **Kafka**, **Spark** Streaming, MLLib, **Python** and utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Used Apache **Spark** Data frames, **Spark**-**SQL**, **Spark** MLLib extensively and developing and designing POC's using **Scala**, **Spark** **SQL** and MLlib libraries.
* Worked on migrating **Map Reduce** programs into **Spark** transformations using **Spark** and **Scala**.
* Designed and developed architecture for data services ecosystem spanning Relational, **NoSQL**, and **Big Data technologies.**
* Used **SQL** Server Integrations Services (**SSIS**) for extraction, transformation, and loading data into target system from multiple sources
* Involved in **Unit Testing** the code and provided the feedback to the developers. Performed **Unit Testing** of the application by using **NUnit.**

**Environment**: Hadoop, Azure, Kafka, Spark, Sqoop, Docker, Swamp, Spark SQL, TDD, Spark-Streaming, Hive, Scala, pig, NoSQL, Impala, Oozie, HBase, Data Lake, Zookeeper.

**Client: Albertson, Boise, ID Oct 2016 – Nov 2018**

**Role:  Data Engineer**

**Responsibilities:**

* Developed **Python** scripts to extract the data from the web server output files to load into **HDFS**.
* Worked on **Cloud Health** tool to generate **AWS** reports and dashboards for cost analysis.
* Written a **python** script which automates to launch the **EMR cluster** and configures the **Hadoop** applications.
* Experience developing **Kafka** producers and **Kafka** Consumers for streaming millions of events per second on streaming data
* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in **Hadoop** and Big Data
* Involved in **SQOOP** implementation which helps in loading data from various RDBMS sources to **Hadoop** systems and vice versa.
* Developed **Map Reduce** programs in Java for parsing the raw data and populating staging Tables.
* Experience in writing **SQOOP** Scripts for importing and exporting data from RDBMS to **HDFS**.
* Worked closely with **AWS** EC2 infrastructure teams to troubleshoot complex issues
* Expertise in writing the **Scala** code using higher order functions for the iterative algorithms in **spark** for performance consideration
* Defined **Kafka** **Zookeeper** offset storage.
* Hands-on use of **Spark** and **Scala** API's to compare the performance of **Spark** with **Hive** and **SQL**, and **Spark** **SQL** to manipulate Data Frames in **Scala**.
* Experience in configuring the **Zookeeper** to coordinate the servers in clusters and to maintain the data consistency which is important for decision making in the process.
* Developed Java **Map Reduce** programs for the analysis of sample log file stored in cluster.
* Saving HUM packet data in **HBASE** for future analytics purpose.
* Involved in Configuring **Hadoop** cluster and load balancing across the nodes.
* Experienced in writing live Real-time Processing using **Spark** Streaming with **Kafka**
* Involved in managing and monitoring **Hadoop** cluster using Cloudera Manager.
* Used **Python** and Shell scripting to build pipelines.
* Developed data pipeline using **sqoop**, HQL, **Spark** and **Kafka** to ingest Enterprise message delivery data into **HDFS**.
* Developed workflow in **Oozie** also in **Airflow** to automate the tasks of loading data into **HDFS** and pre-processing with **Pig** and **Hive**.
* Assisted in creating and maintaining Technical documentation to launching **HADOOP** Clusters and even for executing **Hive** queries and **Pig** **Scripts**.
* Involved in file movements between **HDFS** and **AWS** S3 and extensively worked with S3 bucket in **AWS**.
* Converted all **Hadoop** jobs to run in EMR by configuring the cluster according to the data size.
* Monitor and Troubleshoot **Hadoop** jobs using Yarn Resource Manager and EMR job logs using Genie and **kibana**.

**Environment:** HDFS, Hive, Java, Sqoop, Spark, Yarn, Cloudera Manager, Cloud Health, Splunk, Oracle, Elastic search, Kerberos, Impala, Jira, Confluence, Shell/Perl Scripting, Python, AVRO, Zookeeper, AWS (EC2, S3, EMR, S3, VPC, RDS Lambda, CloudWatch etc), Ranger, Git, Airflow.

**Client: Maisa Solutions Private Limited Hyderabad, India Mar 2015 to July 2016**

**Role: Data Engineer**

**Responsibilities:**

* Using **Sqoop** utility, we transformed the data and populated on to **Hadoop Ecosystem**
* Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.
* Involved in converting **Map Reduce** programs into **Spark** transformations using **Spark** RDD's using **Scala** and **Python**.
* Implemented data streaming capability using **Kafka** and Talend for multiple data sources.
* Worked with multiple storage formats (**Avro**, **Parquet**) and databases (**Hive**, **Impala, Kudu).**
* Worked extensively on **AWS** Components such as Elastic **Map Reduce** (**EMR**)
* Utilized Pandas to create a data frames
* Imported a csv dataset into a **data frame** using pandas.
* Ingested data from varies RDBMS sources.
* Wrote **python** code to manipulate and organize **data frame** such that all attributes in each field were formatted identically.
* Build machine learning models to showcase Big data capabilities using Py**spark** and MLlib.
* Imported data using **Sqoop** to load data from **MySQL** to **HDFS** on regular basis.
* Configured **Zookeeper** to restart the failed jobs without human intervention.
* Working knowledge of cluster security components like **Kerberos, Sentry, SSL/TLS etc.**
* Implemented data streaming capability using **Kafka** and Talend for multiple data sources.
* Capable of using **AWS** utilities such as **EMR, S3** and cloud watch to run and monitor **Hadoop** and **Spark** jobs on **AWS**.
* Migrated **Map reduce** jobs to **Spark** jobs to achieve better performance.
* Implemented **SQOOP** for large dataset transfer between **Hadoop and RDBMS**
* Utilized Matplotlib to graph the manipulated **data frames** for further analysis.
* Graphs provided the data visualization needed to obtain information in a simple form.
* Exported manipulated **data frames** to **Microsoft** **Excel** and utilized its choropleth map feature.
* Created a PowerPoint presentation on the information discovered using data visualization techniques such as bar graphs and choropleth maps

**Environment:** Hadoop, Map Reduce, HDFS, Pig, Hive QL, MySQL, UNIX Shell Scripting, Tableau, Java, Spark, SSIS.

**Client: Ceequence Technologies Hyderabad, India Sep 2013 to Feb 2015**

**Role: Data Engineer**

**Responsibilities:**

* Implemented Avro and parquet data formats for apache Hive computations to handle custom business requirements.
* Designed, implemented and deployed within a customer’s existing Hadoop / Cassandra cluster a series of custom parallel algorithms for various customer defined metrics and unsupervised learning models.
* Installed and configured **Hive, Pig, Sqoop, Flume** and **Oozie** on the **Hadoop cluster.**
* Installed **Oozie workflow** engine to run multiple **Hive** and **Pig Jobs.**
* Developed Simple to complex **Map/reduce** Jobs using **Hive** and **Pig**
* Developed **Map Reduce** Programs for data analysis and data cleaning.
* Extensively used **SSIS** transformations such as Lookup, Derived column, Data conversion, Aggregate, Conditional split, SQL task, Script task and Send Mail task etc.
* Performed data cleansing, enrichment, mapping tasks and automated data validation processes to ensure meaningful and accurate data was reported efficiently.
* Implemented **Apache PIG** scripts to load data from and to store data into Hive.

**Environment:** Hive, Hadoop, Cassandra, Pig, Sqoop, Ooze, Hive, Python, MS Office.