**Professional Summary –**

* 8+ years of professional experience as Data Engineer and combined with Big Data implementations and Spark technologies.
* Experience in Big Data ecosystems using Hadoop, Pig, Hive, HDFS, **MapReduce**, Sqoop, Storm, **Spark, Airflow, Snowflake, Teradata, Flume, Kafka, Yarn, PyCharm SQL server**.
* Hadoop working environments include **Spark, MapReduce, Kafka, Hive, Sqoop,** **HBase,** and **Impala.**
* **Fluent programming experience** with **Scala, Java, Python, SQL, T-SQL, R.**
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, **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** and **Spark YARN.**
* Experience in application of various data sources **like Oracle SE2,** SQL Server, Flat Files and Unstructured files into a data warehouse.
* Experience in analyzing data from Azure data storages using **Databricks** for deriving insights using Spark cluster capabilities.
* Extensive experience in Amazon Web Services (AWS) Cloud services such as **EC2, VPC, S3, IAM, EBS, RDS, ELB, VPC, Route53, OpsWorks, DynamoDB, Autoscaling, CloudFront, CloudTrail, CloudWatch, CloudFormation, Elastic Beanstalk, AWS SNS, AWS SQS, AWS SES, AWS SWF & AWS Direct Connect.**
* Hands on experience in migrating on premise **ETL**s to Google Cloud Platform **(GCP**) using cloud native tools such **as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.**
* Created Automation infrastructure for **Kafka clusters** different instances as per components in cluster using **Terraform** for creating multiple **EC2 instances** & attaching ephemeral or **EBS volumes** as per instance type in different availability zones & multiple regions in AWS.
* Developed custom **RPA bots using Python** and other programming languages to automate data extraction, transformation, and loading (ETL) processes, resulting in increased productivity and reduced costs.
* Implemented **RPA** solutions to streamline financial reporting and analysis, including automating the generation of financial statements, forecasting models, and compliance reports.
* Implemented **AWS X-Ray** allows you to visually detect node and edge latency distribution directly from the service map Tools, **like Splunk, Sumo Logic**, Databrick can be used for log analysis but when comes to distributed tracing within the **AWS, X-Ray** will be provided much better features with service map, traces within depth analysis with minimal configuration with not much maintenance.
* Knowledge of **High Availability** (HA) and **Disaster Recovery** (DR) options in AWS.
* Hands-on experience in Architecting **Legacy Data** Migration projects such as **Teradata** to AWS Redshift migration and from on-premises to A**WS Cloud.**
* Designed, built, and deployed a multitude of applications utilizing almost all AWS Databrick **AWS stack** (Including **EC2, R53, S3, RDS, HSM DynamoDB, SQS, IAM, and EMR.**
* As a **Kubernetes** user in an AWS data engineering role, I have responsibilities which would typically involve designing, deploying, and managing containerized applications.
* Hands on experience in **Kubernetes clusters** for ensuring the scalability, availability, and security of these applications in a highly dynamic and distributed environment.
* Strong hands-on experience with Microservices like Spring IO, Spring Boot in deploying various cloud Infrastructure like AWS and **Spark jobs** on **Databricks.**
* Expertise in configuration and automation using Chef, Chef with **Jenkins, Puppet, Ansible** and **Docker** configuring **Docker Containers** for Branching and deployed using **Elastic Beanstalk**.
* Experience in troubleshooting and resolving architecture problems including database and storage, network, security, and applications.
* Designed and **implemented RPA** solutions to automate data processing tasks, resulting in improved data accuracy and faster processing times.
* **Integrated RPA** solutions with existing data engineering workflows to streamline data processing and enhance efficiency.
* Implemented Copy activity, Custom **Azure Data Factory Pipeline** Activities for **On-cloud ETL** processing and **DWH/BI project** implementation using Azure Data Factory.
* Experience in configuration of report server and report manager for job scheduling, giving permissions to a different level of users in **SQL Server Reporting Services** (SSRS).
* Expert in creating, debugging, configuring, and deploying ETL packages designed MS SQL Server Integration Services **(SSIS**). Configure SQL Azure firewall for a security mechanism.
* Worked in wearing multiple hats**: Azure Architect/System** Engineering, network operations and data engineering.
* used **SAS** with Power BI, connected **Power BI to SAS** data sources, such as **SAS datasets or SAS OLAP cubes**, and then used the data in Power BI for visualization and reporting.
used SAP with Power BI, connected Power BI **to SAP** data sources, such as SAP Business **Warehouse (BW) or SAP HANA**, and then used the data in **Power BI** for visualization and reporting.
* Design & implement migration strategies for traditional systems on Azure (Lift and **shift/Azure** Migrate, other third-party tools).
* Collaborate with application architects on infrastructure as a service (IaaS) application to Platform as a Service (PaaS). **Deploy Azure Resource Manager JSON** Templates from **PowerShell.**

**Technical Skills –**

|  |  |
| --- | --- |
| Big Data Eco Systems | Hadoop, MapReduce, YARN, Kafka, Spark, Airflow, MongoDB, Cassandra, HBase, Storm, Hive, Spark Terraform. |
| Hadoop Distribution | Cloudera Distribution and Horton works. |
| Programming Languages | Scala, JSON, HTML, CSS, JavaScript, SQL, R, Shell Scripting, and Python |
| Database | Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, Snowflake, NoSQL, MongoDB |
| Cloud Platforms | AWS, (EC2, S3, RDS, Red Shift, Glue) Azure, ( ADF, DATA LAKE STORAGE GEN2, DATABRICKS) , GCP |
| Operating Systems | Linux, Unix, Windows |
| Reporting Tools | ETL Tools, Informatica Power Centre, Pentaho, SSIS, SSRS, SAS. |
| Version Control | GIT and Tortoise SVN |
| Analytics Tool | Power BI, Tableau, Data brick |

**Professional Experience –**

**Client** - Huntington Bank, Columbus, OH **Duration** –FEB’2022 - PRESENT

**Role** – Data Engineer-GCP

**Responsibilities** –

* working with product teams to create various store level metrics and supporting data pipelines written in GCP's big data stack.
* By using App teams to collect information from google analytics 360 and built data marts in big query for analytical reporting for the sales and products team.
* building data pipelines in airflow as a service (composer) using various operators
* Building a program with Python dk with Apache beam framework and execute it in Cloud Dataflow to stream pub sub messages into big query tables and deploying streaming maven build cloud dataflow jobs.
* using Python and Apache beam to execute it in cloud Dataflow and to run Data validation jobs between raw source file and big query tables.
* Using Cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query.
* Appling partitioning and clustering for high volume tables on high cardinality fields in Big Query to make queries more efficient.
* loading and transforming large sets of tech structured, semi-structured dataset and analyzed them by running Hive queries.
* Using Cloud Functions to support the data migration from Big Query to the downstream applications.
* Developing scripts using Pyspark to push the data from GCP to the third-party vendors using their API framework.
* Moving data into CP using SQOOP process, using custom hooks for MySQL, using cloud data fusion for moving data from Teradata to GCS.
* Extensively using of cloud shell SDK in GCP to configure/deploy the services like Cloud Dataproc (Managed Hadoop), Google Cloud Storage and Cloud Big query.
* Creating Big Query jobs for loading the data into Big Query tables from data files stored in Google Cloud storage daily.
* Develop reports using Tableau, which keeps track of the dashboards published to Tableau Server, which helps us find potential future clients in the organization.
* Help teams to identify the Big query usage patterns and tuned Big query queries fired from dataflow jobs and more with respect to how app teams can use the its tables for store level attributes.
* Loading data every 15 min on incremental basis to big query raw, Google Data Proc, GCS bucket, hive, Spark, Scala, Python, Gustil and Shell Script.
* Designing star schema in Big Query and architecting various layer of Data Lake.
* Using rest API with Python to ingest Data from and some other site to BIGQUERY.
* Customizing code in python for tagging tables and columns using cloud data catalog and built an application for user provisioning.
* Building a Scala and spark based configurable framework to connect common Data sources like SQL Server, Salesforce, Big query.
* Monitoring Big query, Dataproc and cloud Data flow jobs via Stack driver for all the environment
* Opening SSH tunnel to Google DataProc to access yarn manager to monitor spark jobs.
* Submitting spark jobs using Gsutil GCP and spark submission get it executed in Dataproc cluster.
* Analyzing various types of raw file like] son, Csv, Xml with Python using Pandas, NumPy etc. and Write Scala program for spark transformation in Google DataProc.

**Environment –** MYSQL, MS SQL Server, T: SOL, Oracle, PL/SOL, Google Cloud Platform: CP Cloud Storage, Big Query, Composer, Cloud Dataproc, Cloud SOL, Cloud Functions, Cloud Pub/Sub, Spark, Azure Storage, Azure Database, Azure Data Factory, Azure Analysis Services’/Reporting: Power BI, Data Studio, Tableau Python, Pandas, NumPy, SciPy, Matplotlib, Shell/Bash, C#, R, Go, Visio, Power Point.

**Client** - F P L, Juno Beach, FL **Duration** –OCT’2020- JAN’2022

**Role** – Big Data Engineer(Azure)

**Responsibilities** –

* Involved in gathering requirements, design, implementation, deployment, testing and maintaining of the applications to meet the organization's needs using SCRUM methodology.
* In scrum meetings and coordinated with Business Analysts to understand the business needs and implement the same into a functional design and business rules.
* Understood the business use cases, integration business, write business & technical requirements documents, logic diagrams, process flow charts, and other application related documents.
* Designed and implemented migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party tools. implemented end-to-end data solutions (storage, integration, processing, visualization) in Azure.
* Designed and implemented database solutions in Azure SQL Data Warehouse, Azure SQL, Analysis.
* Implemented Copy activity, Custom Azure Data Factory Pipeline Activities. Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Rarely to build a data pipeline we used to assist a Kubernetes that can be used to manage the deployment and scaling of data pipelines built with tools such as Apache Airflow or Luigi.
* Used Azure Data Factory extensively for ingesting data from disparate source systems and orchestration tool for integrating data from upstream to downstream systems.
* Used and maintained Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
* Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters. Integrated Azure Active Directory authentication to every Cosmos DB request sent and demoed feature to Stakeholders.
* Improved performance by optimizing computing time to process the streaming data and saved cost to the company by optimizing the cluster run time.
* Used Azure Data Factory extensively for ingesting data from disparate source systems.
* Used Pandas in Python for Data Cleansing and validating the source data.
* Designed and developed ETL pipeline in Azure cloud with assistance of Glue jobs which gets customer data from API and processes it to Azure SQL Data Brick.
* Created Databricks Job workflows which extracts data from SQL server and upload the files to sftp using PySpark and python.
* Primarily Involved implement database solutions in Azure SQL Data Warehouse, Azure SQL, storage, and Azure Data Factory, SSIS, PowerShell.
* Created applications to load data from Azure storage blob to Azure SQL, to load from web API to Azure SQL and scheduled web jobs for daily loads.
* Recreated existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database and SQL Data Warehouse environment. experience in DWH/BI project implementation using Azure DF and Databricks. Architect, design and validate Azure infrastructure-as-a-Service (IaaS) environment.
* Developed dashboards and visualizations to help business users analyze data as well as providing data insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS) and Power BI.
* Strong analytical skills, ability to work well in a multi-disciplined team environment and adapt at learning new tools. Responsible for creating Requirements Documentation for various projects.
* Used volatile tables and derived queries for breaking up complex queries into simpler queries.
* Responsible for performance monitoring, resource and priority management, space management, user management, index management, access control, execute disaster recovery procedures.

**Environment -** Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Teradata Utilities, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, Erwin Data Modelling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub.

**Client** - T-Mobile, VA **Duration** – JUL'2018 - SEP’2020

**Role** – Data Engineer – AWS

**Responsibilities** –

* Worked and gathered requirements, analysis of entire system and providing estimation on development, testing efforts.
* Involved in designing different components of system like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
* Hive and spark queries using optimized ways like using window functions, customizing Hadoop shuffle & sort parameter.
* Developed ETL using PySpark. Used both Data frame API and Spark SQL API.
* Used Spark, performed various transformations and actions and the final result data is saved back to HDFS from there to target database Snowflake.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing an storage, Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Knowledge of real time data analytics using Spark Streaming, Kafka, and Flume.
* Configured Spark streaming to get ongoing information from Kafka and store the stream information to HDFS.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Used various spark Transformations and Actions for cleansing the input data.
* Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment. Enforced standards and best practices around data catalog, data governance efforts
* Created DataStage jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel.
* Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc.
* Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Worked in building ETL pipeline for data ingestion, data transformation, data validation on cloud service AWS, working along with data steward under data compliance.
* Worked on scheduling all jobs using Airflow scripts using python added different tasks to DAG, LAMBDA.
* Used Pyspark for extracting, filtering and transforming the Data in data pipelines.
* Skilled in monitoring servers using Nagios, Cloud watch and using ELK Stack Elasticsearch Kibana
* Used Data Build Tool for transformations in ETL process, AWS lambda, AWS SQS
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark databricks cluster.
* Created Unix Shell scripts to automate the data load processes to the target Data Warehouse.
* Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.
* Worked on scheduling all jobs using Airflow scripts using python. Adding different tasks to DAG's and dependencies between the tasks.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS

**Environment -** AWS (EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, Redshift, Cloud formation, CloudWatch, ELK Stack), Jenkins, Ansible, Python, Shell Scripting, PowerShell, GIT, Microservice, Jira, JBOSS, Bamboo, Kubernetes, Docker, Web Logic, Maven, Web sphere, Unix/Linux, Nagios, Splunk, AWS Glue, Data Brick

**Client** - FEPOC, VA. **Duration** – JAN'2017- JUN'2018

**Role** – Kafka Hadoop Data Engineer

**Responsibilities** –

* Involved in importing the real time data to Hadoop using Kafka and implemented the Oozie job for daily data.
* Created stored procedures in MySQL Server to perform result-oriented tasks.
* Managed and reviewed Hadoop Log files as a part of administration for troubleshooting purposes. Communicate and escalate issues appropriately.
* Responsible for designing highly scalable big data clusters to support storage and computation across varied big data clusters - Hadoop, Cassandra, MongoDB & Elastic Search.
* Responsible for installing, configuring, supporting, and managing of Installed Kerberos secured Kafka cluster wif no encryption on POC a O Created NoSQL solution for a legacy RDBMS Using Kafka, Spark, Hadoop Clusters.
* Kafka ACL's HBase indexer for ingestion SOL and HBase for and real-time querying
* Experienced in Administration, Installing, Upgrading and Managing distributions of Hadoop clusters wif MaR 5.1 on a cluster of 200+ nodes in different environments such as Development, Test and Production (Operational & Analytics) environments.
* Troubleshooting issues in teh execution of MapReduce jobs by inspecting and reviewing log files.
* Worked on implementing the NOSQL database Cassandra cluster.
* Extensively worked on Elastic search querying and indexing to retrieve teh documents in high speeds.
* Installed, configured, and maintained several Hadoop clusters which includes HDFS, YARN, Hive, HBase, Knox, Kafka, Oozie, Ranger, Atlas, Infra Sol, Zookeeper, and Nifi in Kerber zed environments.
* Installed and configured Hadoop, Map Reduce, HDFS (Hadoop Distributed File System), developed multiple Map Reduce
* jobs in java for data cleaning.
* Experience in managing teh Hadoop cluster wif IBM Big Insights, Hortonworks Distribution Platform. Regular Maintenance of Commissioned/decommission nodes as disk failures occur using Map File
* Experience in managing teh Hadoop cluster wif IBM Big Insights, Hortonworks Distribution Platform.
* Responsible for installing, configuring, supporting, and managing Cloudera Hadoop Clusters.
* Worked on installing cluster, commissioning & decommissioning of Data Nodes, Name Node recovery, capacity planning and slots configuration in Map Control System (MCS).
* Experience in innovative, and where possible, automated approaches for system administration tasks.
* Worked on setting up high availability for major production cluster and designed automatic failover control using zookeeper and quorum journal nodes.
* Used Spark Streaming to divide streaming data into batches as an input to spark engine for batch processing. Mentored
* EQM team for creating Hive queries to test use cases.
* Sqoop configuration of JDBC drivers for respective relational databases, controlling parallelism, controlling Distich ache, controlling import process, compression codecs, importing data to hive, Base, incremental imports, configure saved jobs.

**Environment -** Hadoop, Python, HDFS, Hive, Scala, MapReduce, Agile, Cassandra, Kafka, Storm, AWS, YARN, Spark, ETL, Teradata, NoSQL, Java, Cassandra, Talend, LINUX, Kibana, HBase.

**Client** - Genems Systems, Hyderabad. **Duration**–JAN’2015- AUG'2016

**ROLE**: BI Data Analyst

**Responsibilities –**

* Analyze functional and non-functional categorized data elements for data profiling and mapping from source to target. Data environment. Developed working documents to support findings and assign specific tasks.
* Involved with data profiling for multiple sources and answered complex business questions by providing data to business users.
* Worked with data investigation, discovery, and mapping tools to scan every single data record from many sources.
* Data mining on Claims data using very complex SQL queries and discovered claims pattern. Created DML code and statements for underlying & impacting databases.
* Used ETL methodology for supporting data extraction, transformations and loading processing, in a complex EDW using Informatica.
* Worked on Data reconciliation between integrated systems.
* Metrics reporting, data mining and trends in helpdesk environment using Access.
* Written complex SQL queries for validating the data against different kinds of reports generated by Business Objects XIR2.
* Used MS Access to pull the data from various databases and integrate the data.
* Involved in Normalization and De-Normalization of existing tables for faster query retrieval. Implemented advanced SQL queries using DML and DDL statements to extract large quantities of data from multiple data points on MS SQL and Teradata SQL Assistant
* Performed data analysis and data profiling using complex SQL on various sources systems including Oracle and Teradata.
* Assisted with data mapping activities to ensure accurate and consistent movement of data from source systems to a Teradata target database while complying with enterprise data standards.
* Used Microsoft SQL Server Integration Services (SSIS) to extract, transform, and load data from various sources such as Excel and flat files to a Microsoft SQL Server database.
* Worked in importing and cleansing data from various sources like Teradata, Oracle, flat files, SQL Server 2005 with high volume data Worked with Excel Pivot tables.
* Created and Configured OLAP Cubes (Star Schema and Snow Flex Schema) using SQL Server Analysis Services (SSAS). Developed data visualizations and dashboards using Power Bl and Tableau.
* Create and Monitor workflows using workflow designer and workflow monitor.
* Performing data management projects and fulfilling ad-hoc requests according to user specifications by utilizing data management software programs and tools like Perl, Toad, MS Access, Excel, and SQL
* Written SQL scripts to test the mappings and Developed Traceability Matrix of Business Requirements mapped to Test. Scripts to ensure any Change Control in requirements leads to test case update.
* Created SSIS packages to implement error/failure handling Designed and implemented SSIS packages that can be easily deployed to different environments (Development, Testing, and Production).
* Worked extensively with SSAS, including managing storage, partitions, and aggregations, as well as developing queries with MDX and data mining models.
* Created reports using MDX and SQL, drawing on knowledge of multidimensional databases and relational databases.
* Wrote calculated columns, Measures query's using DAX in Power BI to show good data analysis techniques, Involved in scorecard and KPI Reporting.
* Designed Tableau and Power BI dashboard which give a visual representation for the users on MTD sales growth, Visual analytics on injuries, dealer location map, various visualization techniques in Tableau like Scatter Plots, Histogram and Heat Maps.
* Involved in extensive DATA validation by writing several complex SQL queries and involved in back-end testing and worked with data quality issues.
* Identified & recorded defects with required information for issues to be reproduced by development team. Flexible to work late hours to coordinate with offshore team.

**Environment**: MS SQL Server 2012/ 2-14/ 2017, Microsoft Management Studio, Microsoft Visual Studio, MS SOL Server Integration Services (SSIS) 2012/2008R2/2008, SQL Server Reporting Services (SSRS), Power BI, Tableau, T-SQL, MDX, OLAP Cubes, DAX, KPI.

**Education Details:**

**Undergraduate**

Bachelors in computer science - Jawaharlal Nehru Technological University Konda Gattu. (JNTUK)

**Subjects:** Advanced data structures, DBMS, Cloud computing, Algorithms Design Python, Java etc.