**Ankitha Punna**

**Sr. Big Data Architect /Cloud Engineer**

Desk : **+1(479)-437-5840**

Email : [sales24@kteksoft.com](mailto:sales24@kteksoft.com)

**PROFESSIONAL SUMMARY:**

* 10 years of experience in SDLC with a primary focus on cutting-edge Big Data Technologies, including Spark, Scala, Spark MLlib, Hadoop, Tableau, and Cassandra.
* Proficient in Big Data and Data Warehouse Architecture, specializing in designing Star Schema, Snowflake Schema, Fact and Dimensional Tables, and both Physical and Logical Data Modeling using Erwin and ER Studio.
* Skilled in architecting, designing, and developing Big Data Solutions, encompassing the establishment of Big Data roadmaps, infrastructure setup, and team building to provide comprehensive Big Data services.
* Successfully led the architecture and implementation of a Portfolio Recommendation Analytics Engine using Hadoop MR, Oozie, Spark SQL, Spark MLlib, and Cassandra.
* Possess a deep understanding of Hadoop architecture and the underlying framework, with a strong emphasis on storage management.
* Extensive expertise in data modeling, data architecture, solution architecture, data warehousing, business intelligence, and master data management (MDM) concepts.
* Proficient in architecting Big Data solutions, encompassing data ingestion and storage.
* Hands-on experience with NoSQL databases such as Hbase, Cassandra, and MongoDB, along with expertise in database performance tuning and data modeling.
* Skilled in designing Extract, Transform, Load (ETL) environments using Informatica Power Center.
* Proficient in writing Stored Procedures, Functions, and Triggers using PL/SQL.
* Experienced in integrating various data sources, including SQL Server, Oracle, Sybase, ODBC connectors, and Flat Files.
* Accomplished in handling large volumes of data in and out of Teradata and Big Data environments.
* Proficient in developing Big Data projects using Hadoop, Hive, HDP, Pig, Flume, Storm, and Map Reduce open-source tools.
* Expertise in architecting, solutioning, and modeling Data Integrity (DI) Platforms using Sqoop, Flume, Kafka, Spark Streaming, Spark MLlib, and Cassandra.
* Strong experience in migrating data warehouses and databases to Hadoop and NoSQL platforms.
* Strong expertise in Amazon AWS services, including EC2, Dynamo DB, S3, Kinesis, and more.
* Extensive experience in data analysis, design, and modeling using tools like ErWin.
* Expertise in Big Data architecture, encompassing Hadoop (Azure, Hortonworks, Cloudera), distributed systems, MongoDB, and NoSQL.
* Proficient in using Teradata BTEQ, FastLoad, MultiLoad, and FastExport utilities.
* Hands-on experience in Hadoop and Big Data technologies for data storage, querying, processing, and analysis.
* Experienced in utilizing various Hadoop infrastructures, such as MapReduce, Hive, Sqoop, and Oozie.
* Expert in Amazon EMR, Spark, Kinesis, S3, Boto3, Beanstalk, ECS, CloudWatch, Lambda, ELB, VPC, Elastic Cache, Dynamo DB, Redshift, RDS, Athena, Zeppelin, and Airflow.
* Proficient in testing data in HDFS and Hive for transactional data.
* Strong technical and analytical skills with a clear understanding of ER modeling for OLTP and dimension modeling for OLAP.
* Extensive experience working with databases like Oracle 12C/11g/10g/9i, DB2, SQL Server 2008, and MySQL, with proficiency in writing complex SQL queries.
* Experienced in database tools such as SQL Navigator and TOAD.
* Skilled in optimizing and improving the performance of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, Pair RDDs, and Spark YARN.
* Knowledgeable in programming with Resilient Distributed Datasets (RDDs).
* Experienced in using Flume for transferring log data files to Hadoop Distributed File System (HDFS).
* Familiar with job workflow scheduling and monitoring tools like Oozie and Zookeeper.
* Proficient in Shell programming.
* Knowledgeable in configuring and managing Cloudera's Hadoop platform, including CDH3 and CDH4 clusters.
* Familiar with the architecture and functionality of NoSQL databases like Cassandra and MongoDB.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Hadoop Components / Big Data** | **HDFS, Hue, MapReduce, PIG, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, pyspark**  **Airflow, Kafka Snowflake,** |
| **Languages** | **Scala, SQL, Python, Hive QL, KSQL. Boto3** |
| **IDE Tools** | **Eclipse, IntelliJ, pycharm, VSCode.** |
| **Cloud platform** | **AWS, Azure** |
| **Reporting and ETL Tools** | **Tableau, Power BI, Talend, Pentaho, AWS GLUE.** |
| **Databases** | **Oracle 19, SQL Server, MySQL, Druid, MS Access, NoSQL Database (HBase, Cassandra, Mongo DB), T-SQL** |
| **Big Data Technologies** | **Hadoop, HDFS, Hive, Pig, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Flink, Flume, Airflow, Informatica, Snowflake, DataBricks, Kafka, Cloudera, RabbitMQ.** |
| **Data Analysis Libraries:** | **Pandas, NumPy, SciPy, Scikit-learn, NLTK, Plotly, Matplotlib** |
| **BI Tools:** | **Alteryx, Tableau Power BI, Sisense, Looker.** |
| **Containerization** | **Docker, Kubernetes** |
| **CI/CD Tools** | **Jenkins, Bamboo, GitLab** |
| **Operating Systems** | **UNIX, LINUX, Ubuntu, CentOS.** |
| **Software Methodologies** | **Agile, Scrum, Waterfall** |
| **Reporting Tools** | **PowerBI, Qlikview, Tableau, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5. X/ 6.x, Cognos7.0/6.0.** |
| **Frameworks** | **Jingo, Flask, Fast API.** |

**PROFESSIONAL EXPERIENCE:**

**PG&E, CA Feb 2021 - Present**

**Sr. Big Data Architect**

**Responsibilities:**

* Implemented solutions for ingesting data from various sources and processing the Data-at Rest utilizing.
* Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, Hive.
* Implementation of Big Data ecosystem (Hive, Impala, Sqoop, Flume, Spark, Lambda) with Cloud Architecture.
* Designed and deployed full SDLC of AWS Hadoop cluster based on client's business need.
* Experience on BI reporting with at Scale OLAP for Big Data.
* Developed Complex ETL code through Data manager to design BI related Cubes for data analysis at corporate level.
* Unified data lake architecture integrating various data sources on Hadoop architecture.
* Used Sqoop to import the data from RDMS to Hadoop Distributed File System (HDFS).
* Involved in loading and transforming large sets of data and analyzed them by running Hive queries and Pig scripts.
* Developed software routines in R, Spark, SQL to automate large datasets calculation and aggregation.
* Integrated NoSQL database like Base with Map Reduce to move bulk amount of data into HBase.
* Redesigned the existing Informatica ETL mappings & workflows using Spark SQL.
* Loaded and transformed large sets of structured, semi structured and unstructured data using Hadoop/Big Data concepts.
* Ingest data into Hadoop/Hive/HDFS from different data sources.
* Writing Scale code to run SPARK jobs in Hadoop HDFS cluster.
* Define and manage the architecture and life cycle of Hadoop and SPARK projects
* Designed and Developed Real time Stream processing Application using Spark, Kafka, Scale and Hive to perform Streaming ETL and apply Machine Learning.
* Designed the data processing approach within Hadoop using Pig.
* Identify query duplication, complexity and dependency to minimize migration efforts Technology stack: Oracle, Horton works HDP cluster, at unity Visibility, Cloudera Navigator Optimizer, AWS Cloud and Dynamo DB.
* Experience in AWS, implementing solutions using services like (EC2, S3, RDS, Redshift, VPC)
* Worked using Apache Hadoop ecosystem components like HDFS, Hive, Sqoop, Pig, and Map Reduce.
* Lead architecture and design of data processing, warehousing and analytics initiatives.
* Worked with AWS to implement the client-side encryption as Dynamo DB does not support at-rest encryption at this time.
* Exploring with Spark for improving the performance and optimization of the existing algorithms in
* Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Used Data Frame API in Scale for converting the distributed collection of data organized into named columns.
* Performed data profiling and transformation on the raw data using Pig, Python.
* Experienced with batch processing of data sources using Apache Spark.
* Developing predictive analytics using Apache Spark Scale APIs.
* Involved in working of big data analysis using Pig and User defined functions (UDF).
* Created Hive External tables and loaded the data into tables and query data using HQL.
* Used Sqoop to efficiently transfer data between databases and HDFS and used Flume to stream the log data from servers.
* Implement enterprise-grade platform (mark logic) for ETL from mainframe to NOSQL (Cassandra).
* Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format.
* Developed Spark streaming application to pull data from cloud to hive table.
* Implemented Spark Graph application to analyze guest behavior for data science segments.
* Enhancements to traditional data warehouse based on STAR schema, update data models, perform Data Analytics and Reporting using Tableau.
* Experience in different Hadoop distributions like Cloudera (CDH3 & CDH4) and Horton Works Distributions (HDP) and Map.
* Experience in integrating oozy logs to kabana dashboard.
* Extracted the data from MySQL, AWS Redshift into HDFS using Sqoop.
* Developed Spark code using Scale and Spark-SQL for faster testing and data processing.
* Managed data extraction, transformation, and loading (ETL) processes using Iceberg to ensure data quality and accuracy.
* Collaborated with cross-functional teams to gather and analyze data requirements for various projects.
* Developed and maintained data pipelines and workflows using Iceberg for efficient data processing.
* Conducted data analysis using Iceberg to identify trends, patterns, and insights.
* Created customized reports and dashboards using Iceberg for stakeholders to make informed decisions.
* Implemented a proof of concept deploying this product in Amazon Web Services AWS. Designed, implemented, and managed cloud-based solutions on the Google Cloud Platform (GCP). This involved leveraging GCP services such as Compute Engine, Kubernetes Engine, and Cloud Functions to optimize infrastructure and achieve scalability.
* Developed and maintained infrastructure as code using tools like Terraforming, enabling efficient deployment and scaling of cloud resources.
* Implemented security and compliance measures, including IAM policies and VPC configurations, to protect sensitive data and applications.
* Automated routine tasks and processes using GCP's tools like Cloud Scheduler and Cloud Monitoring, streamlining operations and improving reliability.
* Troubleshooter and resolved technical issues related to GCP services, ensuring minimal downtime and uninterrupted service.
* Regularly updated knowledge of GCP services and features to leverage the latest cloud innovations for improved project outcomes.

**Environment: Agile, AWS, Cloud Watch, Cloud Trial, Cloud Formation, Athena, AWS Glue, Mongo DB, Docker, ETL, GCP, snowflake, Map Reduce, Sqoop, API, Oozie, Looker, Jiri, Zookeeper, Pentaho, GIT, Python, Hadoop, Map Reduce, HDFS, Hive, Presto, Apache Flink, T-SQL, Python, Druid, DB2, Streaming, SQL, Amazon RDS, Amazon EC2, S3, Cloud Watch, Spark, Scale, Get, Kafka, Redshift, Dynamo DB, PostgreSQL, ORMB and star schema, Data Modeling.**

**SMBC Bank, New York City, NY**  **Nov 2019 – Jan 2021**

**Lead Data Engineer**

**Responsibilities:**

* Led the implementation of a robust CI/CD process using GitLab, Python, and Shell scripting to automate routine tasks. This included the synchronization of installers, configuration modules, packages, and application requirements.
* Constructed AWS data pipelines using a suite of AWS services, including VPC, EC2, S3, Auto Scaling Groups (ASG), EBS, Snowflake, IAM, CloudFormation, Route 53, CloudWatch, CloudFront, and CloudTrail.
* Created and configured elastic load balancers and auto scaling groups to distribute traffic, ensuring a cost-efficient, fault-tolerant, and highly available environment.
* Managed metadata alongside data in AWS Data Lake, using complex functions like AWS Lambda and AWS Glue for efficient data tracking.
* Developed software systems with scientific analysis and optimal algorithms, including sessionization for generating statistics in web site analyzers.
* Designed and developed RESTful and SOAP APIs using Swagger and conducted comprehensive testing using Postman.
* Demonstrated proficiency with Redshift Database, handling ETL data pipelines from AWS Aurora (MySQL Engine) to Redshift.
* Implemented data organization strategies like Distkey, Sort key, and groupings for improved query performance.
* Managed user-level and group-level permissions within the Redshift schema to enhance security.
* Designed and constructed AWS Data pipelines utilizing various AWS resources, including AWS API Gateway, AWS Lambda, Snowflake, DynamoDB, and AWS S3, to facilitate data retrieval and transformation.
* Built end-to-end ETL pipelines from AWS S3 to DynamoDB and Snowflake Data Warehouse, facilitating analytical queries and cloud data processing.
* Transformed data into different formats to meet user and business requirements from diverse sources, including Snowflake and unstructured data.
* Documented all changes made across systems and components using Confluence and Atlassian Jira. This encompassed technical changes, infrastructure changes, and business process changes.
* Comprehensively maintained post-release documentation, including known issues from production implementation and deferred defects.
* Proficiently managed Informatica Cloud Services and Informatica Power Center for ETL strategies and mappings.
* Set up Secure Agents and data connectors to process unstructured, semi-structured, and structured data for various applications.
* Led the successful migration of critical business applications to Google Cloud Platform (GCP), resulting in a 30% reduction in operational costs and improved system performance.
* Designed and implemented scalable cloud architectures, including Virtual Private Cloud (VPC), Kubernetes Engine, and Cloud Functions, to ensure high availability and security of cloud-based services.
* Collaborated with cross-functional teams to define cloud strategy, optimize resources, and meet business objectives, ensuring seamless integration of cloud services.
* Implemented automated deployment and CI/CD pipelines using tools like Cloud Build and Jenkins, reducing deployment times by 40%.
* Monitored and maintained cloud infrastructure using GCP monitoring and logging tools, proactively addressing performance issues and security vulnerabilities.
* Ensured compliance with industry-specific regulations and best practices, conducting regular security audits and vulnerability assessments.
* Provided training and guidance to the IT team to foster GCP expertise and promote best practices for cloud management.
* Documented cloud architecture, configurations, and policies, facilitating knowledge transfer and future scalability.
* Managed data governance and data quality initiatives using IceBerg.
* Developed and implemented data models and schemas on IceBerg for efficient data storage and querying.
* Created and maintained ETL (Extract, Transform, and Load) processes to ingest and process data into IceBerg.
* Collaborated with cross-functional teams to define data retention policies and data access controls in IceBerg.
* Worked on connecting and transferring data from various sources to IceBerg for centralized storage and analysis.
* Tuned IceBerg tables and queries to enhance query performance and reduce latency.
* Implemented best practices for partitioning and clustering data in IceBerg to improve query efficiency.

**Environment:** **Agile, Azure, Hadoop, MapReduce, Oozie, Pig, Zookeeper, HDFS, Hive, Docker, Druid, GIT, ETL, Python, Glue, JIRA, NoSQL, Mongo DB, Kubernetes, Cassandra, Tableau, Power BI, ORMB and starschema.**

**Change Healthcare, Chicago,IL Feb 2018 – Oct 2019**

**Data Engineer/Cloud Engineer**

**Responsibilities:**

* Involved in SDLC Requirements gathering, Analysis, Design, Development, and Testing of applications using Agile Methodology (Scrum).
* Worked on creating a data pipeline of gathering, cleaning, and optimizing data using Hive and Spark.
* Responsible for building scalable distributed data solutions using an EMR cluster environment with Amazon EMR.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka and Persists into HDFS.
* Designed and developed data integration programs in a Hadoop environment with NoSQL data store Cassandra for data access and analysis.
* Imported and Exported Data from Different Relational Data Sources like SQL Server to HDFS using Sqoop.
* Strong knowledge and experience in creating Jenkins CI/CD pipelines and good experience in full automation via Jenkins, Git, and Docker.
* Developed Spark SQL scripts using Python for faster data processing.
* Implemented a 'serverless' architecture using API Gateway, Lambda, and Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets.
* Worked huge datasets stored in AWS S3 buckets, and used spark data frames to perform preprocessing in Glue.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Updated Python scripts to match training data with our database stored in AWS Cloud Search, so that we would be able to assign each document a response label for further classification.
* Developed spark workflows using Scala to pull the data from AWS and apply transformations to the data.
* Developed MapReduce/Spark Python modules for machine learning and predictive analytics in Hadoop on AWS.
* Worked on Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie.
* Developed Spark Applications by using Python and Implemented Apache Spark data processing Project to handle data from various RDBMS and Streaming sources.
* Worked on Migrating MapReduce programs into Spark transformations using Spark and Scala.
* Extracted the data from HDFS using Hive and performed data analysis using Spark with Scala, PySpark, and Redshift for feature selection and created nonparametric models in Spark.
* Also worked on RDS databases like MySQL servers and NOSQL databases like MongoDB, HBase.
* Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using R and Tableau.
* Developed Tableau visualizations and dashboards using Tableau Desktop.

**Environment:** **Python, R, SQL, Hive, Spark, AWS, Hadoop, NoSQL, Cassandra, SQL Server AWS, HDFS, PySpark, Tableau, Mongo DB, Postgres SQL, Redshift, HBase, Sqoop, Airflow, Oozie.**

**HSBC Bank, Hyderabad, India Jun 2015- Dec 2017**

**Data Engineer/Cloud Engineer**

**Responsibilities:**

* Utilized Oozie workflow engine to manage Hadoop jobs and automate tasks in various Hadoop components like Java MapReduce Hive, Pig, and Sqoop.
* Collaborated with cross-functional teams, including data scientists, software engineers, and business stakeholders, to deploy and monitor machine learning models in the cloud.
* Worked with Agile and Scrum methodologies.
* Analyzed critical datasets using Cloudera, HDFS, HBase, MapReduce, Hive, Pig, Sqoop, Spark, and ZooKeeper.
* Implemented NoSQL data modeling, tuning, disaster recovery, and backup using MongoDB.
* Created HBase tables and loaded large sets of semi-structured data from different sources.
* Developed solutions using Java, Spring, and Hadoop technologies, including riDFS, Hive, and MapReduce.
* Wrote Hive and Pig scripts for ETL tasks, transformations, event joins, filtering, and pre-aggregations.
* Designed and implemented data pipelines using Flume, Sqoop, Pig, and MapReduce to ingest customer behavioral data.
* Developed Spark code using Scala and Spark-SQL for faster testing and data processing.
* Used Spark API on Cloudera Hadoop YARN to analyze data from Hive.
* Moved log files from various sources to HDFS for processing using Flume.
* Developed Java code for generating, comparing, and merging AVRO schema files.
* Read and wrote multiple data formats (JSON, ORC, Parquet) on HDFS using PySpark.
* Prepared validation report queries and shared results with business users after ETL runs.
* Installed and configured Pentaho BI server for ETL and reporting services.
* Worked with cloud-based technologies like Redshift, S3, AWS, EC2 Machine, Glue, and loaded data from Oracle financials and Redshift database.
* Performed various testing activities, including functional, integration, regression, smoke, and performance testing, on Hadoop MapReduce developed in Python, Pig, and Hive.
* Designed and implemented data models, routes, and controllers using Python and FastAPI.
* Integrated FastAPI with databases such as PostgreSQL, MongoDB, and MySQL.
* Developed a common Flink module for serializing and deserializing AVRO data.
* Experience with batch processing and streaming frameworks like Kafka and Apache Flink.
* Developed complex APIs using FastAPI, adhering to best practices and project requirements.
* Designed and implemented data warehouses on GCP using BigQuery.
* Deployed FastAPI applications using Docker and Kubernetes for scalability.
* Utilized GCP services for data processing, storage, and analysis (BigQuery, Cloud Storage, Dataflow, Dataproc, Pub/Sub, Cloud SQL).
* Implemented machine learning algorithms using Spark and Python.
* Experience with CI/CD principles and tools like Jenkins and Git for building and deploying ML models.
* Proficient in performance optimization and scaling of ML pipelines in distributed systems.
* Worked on AWS MLOPS services (Sagemaker, ECR, Codebuild, Docker, Kubernetes), cloud functions, Terraform, and Git CI/CD.
* Implemented Spark using Scala and SparkCore, Spark Streaming, and SparkSQL API for faster data processing.
* Configured monitors, alarms, notifications, and logs using Cloudwatch for Lambda functions, Glue Jobs, and EC2 hosts.
* Created Pyspark framework for transferring data from DB2 to Amazon S3.
* Imported data from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
* Integrated Apache Storm with Kafka for web analytics and clickstream data processing.

**Environment: Agile, AWS, S3, EC2, S3, RedShift, RDS, Route53, EMR, Elastic Search, Cloudera, HDFS, HBase, MapReduce, Hive, Pig, Sqoop, Flink, Spark and Zookeeper, Fast API, Mongo DB, Cassandra, GIT, Spark SQL, Spark RDD, Python, Tableau.**

**Zero Touch, Hyderabad, India**  **May 2013 to May 2015**

**Cloud Engineer**

**Responsibilities:**

* Worked with Hadoop ecosystem components like MapReduce, HDFS, HBase, Zookeeper, Hive, Pig, Spark, Pyspark, Oozie, Storm, Yarn, and Snowflake.
* Utilized NoSQL databases, specifically HBase, to create tables and load large sets of semi-structured data from different sources.
* Performed ETL (Extract, Transform, and Load) processes using Python-SQL Server pipelines/framework for data analytics and visualization using libraries such as NumPy, SciPy, Pandas, and MATLAB stack.
* Transferred tables from Teradata to Hadoop using Sqoop.
* Followed the waterfall methodology.
* Developed custom reports and modified queries and reports in Power BI desktop and SSRS (SQL Server Reporting Services).
* Created data engineering and ETL Python scripts for ingestion pipelines running on AWS infrastructure (EMR, S3, Glue, Lambda).
* Managed data import from various sources, performed transformations using Hive and MapReduce, loaded data into HDFS, and extracted data from MySQL into HDFS using Sqoop.
* Collaborated with application teams to install operating systems, update Hadoop, and perform version upgrades.
* Analyzed Hadoop cluster and utilized big data analytic tools like MapReduce, Hive, and Spark.
* Loaded data from LINUX file systems, servers, and Java web services using Kafka Producers.
* Migrated complex MapReduce programs to Spark RDD transformations and actions.
* Used Spark Streaming to receive real-time data from Kafka, storing the stream data to HDFS using Scala, and NoSQL databases like HBase and Cassandra.
* Worked on user profile and unstructured data storage using Java and MongoDB, developing core modules for user analytics and prediction engines using Scala.
* Integrated with U layer using HIML, Ajax, and JavaScript.
* Automated jobs, including data pulling from storage, loading data into MySQL, and utilizing ShellScripts.
* Automated end-to-end data processing, including pulling data from various sources, pushing result datasets to Hadoop Distributed File System, and running MapReduce jobs and PIG/Hive using Kettle and Oozie (Workflow management).
* Collaborated with regulatory delivery leads to ensure robustness in prop trading control frameworks using Hadoop, Python Jupyter Notebook, Hive, and NoSQL.
* Created Apache Airflow pipelines using Python to schedule daily/weekly jobs.

**Environment:** **Hadoop, Spark, HBase, HDFS, Map Reduce, AWS, Teradata, SQL, Cloudera, Pig, Latin, Sqoop, Hive, Pig, MySQL, Oozie, Flume, Informatica, Zookeeper, R, Python, ETL, Tableau.**

**EDUCATIONAL DETAILS:**

* **Bachelor’s in Computer Science, Osmania University, 2013.**