**Kishan Contact: +1 972 8538231**

**Sr Data Engineer Email: krishna@adven-it.com**

**PROFESSIONAL SUMMARY**

* Overall, **9+ years** of technical IT experience in Design, Development, and Maintenance of enterprise analytical solutions using big data technologies. Proven data and business expertise in Retail, Finance, and Healthcare domains.
* A result-oriented professional with experience in Creating Data Mapping Documents, Writing Functional Specifications and Queries, Normalizing Data from 1NF to 3NF/4NF. Requirements gathering, System & Data Analysis, Requirement Analysis, Data Architecture, Database Design, Database Modeling, Development, Implementation, and Maintenance of **OLTP** and **OLAP** Databases.
* Strong knowledge and experience on Data Analysis, Data Lineage, Big Data pipelines, Data quality, Data Reconciliation, Data transformation rules, Data flow diagram including Data replication, Data integration and Data orchestration tools.
* Strong Experience in working with **ETL Informatica** ( 10.4/10.1/9.6/8.6/7.1.3 ) which includes components **Informatica PowerCenter Designer, Workflow manager, Workflow monitor, Informatica server** and **Repository Manager**.
* Vast experience in designing, creating, testing, and maintaining the complete data management from **Data Ingestion**, **Data Curation** and **Data Provision** with in-depth knowledge in **Spark API's** like **Spark Framework-SQL**, **DSL**, Streaming, by working on different file formats like **parquet**, **JSON**, and performance tuning of spark applications from various aspects.
* Provided full life cycle support to logical/physical database design, schema management, and deployment. Adept at database deployment phase with strict configuration management and controlled coordination with different teams.
* Experience in writing code in **R** and **Python** to manipulate data for data loads, extracts, statistical analysis, modeling, and data munging.
* Utilized **Kubernetes** and **Docker** for the runtime environment for the **CI/CD** system to build, test, and deploy.
* Utilized analytical applications like **R** and **Python** to identify trends and relationships between different pieces of data, draw appropriate conclusions and translate analytical findings into risk management and marketing strategies that drive value.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie. Experienced with using most common Operators in **Airflow** - **Python Operator, Bash Operator, Google Cloud Storage Download Operator, Google Cloud Storage Object Sensor, GoogleCloudStorageToS3Operator.**
* Extensive hands-on experience in using distributed computing architectures such as AWS products (e.g., EC2, Redshift, EMR, and Elastic search), Hadoop, Python, Spark, and effective use of Azure SQL Database, MapReduce, Hive, SQL, and PySpark to solve big data type problems.
* Strong experience in Microsoft Azure Machine Learning Studio for data import, export, data preparation.
* Proficient in Statistical Methodologies including Hypothetical Testing, **ANOVA**, Time Series, Principal Component Analysis.
* Good understanding of Data Modeling (Dimensional and Relational) concepts like Star-Schema Modeling, **Snowflake** Schema Modeling, Fact and Dimension Tables.
* Strong Experience in working with Linux/Unix environments, writing Shell Scripts.
* Good at conceptualizing and building solutions quickly and recently developed a Data Lake using sub-pub Architecture.
* Hands on experience with **Guidewire** Policy Center and Claim Center (Worker's Compensation Claims, General Claims) upgradation to v9.
* Developed a pipeline using Python and Kafka to load data from a server to Hive with automatic ingestions and quality audits of the data to the RAW layer of the Data Lake.
* Installed both **Cloudera** (CDH4) and **Hortonworks** (HDP1.3-2.1) **Hadoop clusters** on **EC2**, **Ubuntu 12.04**, **CentOS 6.5** on platforms ranging from 10-100 nodes.
* Architected complete scalable data pipelines, a data warehouse for optimized data ingestion.
* Collaborated with data scientists and architects on several projects to create a data mart as per requirement.
* Conducted complex data analysis and reported the results. Constructed data staging layers and fast real-time systems to feed BI applications and machine learning algorithms.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Data Technologies** | AWS, S3, Lambda, Triggers, Glue, EMR, Kinesis, Redshift, Hadoop, HDFS, Hive, MapReduce, Pig, Flume, Oozie, HBase, Spark. |
| **Programming** | Python, PySpark, Scala, Java, C, C++, Shell script, Perl script, SQL. |
| **AWS Services** | Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sagemaker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation. |
| **Databases** | MySQL, SQL/PL-SQL, MS-SQL Server, Oracle, Teradata 12.0/13.0. |
| **Python Libraries/Packages** | NumPy, SciPy, Boto, Pickle, PySide, PyTables, Data Frames, Pandas, Matplotlib, SQLAlchemy, HTTPLib2, Urllib2, Beautiful Soup, Py Query. |
| **ETL Tools** | Cassandra, Hbase |
| **Visualization/ Reporting** | Tableau, ggplot2, matplotlib, SSRS and Power BI. |
| **Tools** | PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Eclipse, Postman |

**EXPERIENCE**

**UPMC Health Plan| Pittsburg, PA September 2020 - Current  
Sr Data Engineer**

**Responsibilities**:

* Worked with the Business team for gathering the requirements and also help them with their test cases.
* Played a vital role in design and development for building the common architecture for retail data across geographies.
* Worked on design and developing 5 different flows Point of sales, Store traffic, Labor, Customers Survey, and Audit data.
* Developed a common framework using spark to ingest data from different data sources (Teradata to S3 and S3 to Snowflake etc.,).
* Developed reusable spark scripts and functions for data processing that can be leveraged in different data pipelines.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Worked on ingesting the real-time data using Kafka
* Designed, developed Azure (AAS & SSAS) cubes for data visualization.
* Used Sqoop to ingest the data from the Oracle database and store them on S3.
* Worked on ingesting data from JSON, CSV files using spark and EMR and store the output data in Parquet file format on S3.
* Built ETL pipelines on Snowflake and the data products are used by stakeholders for querying and serve as backend objects for visualizations
* Worked on performance tuning of spark jobs by adjusting the memory parameters and the cluster configuration.
* Excellent knowledge of AWS services (S3, EMR, Athena, EC2), Snowflake, and Big Data technologies.
* Providing knowledge transition to support team
* Used Airflow for scheduling and orchestration of the data pipelines.

**Environment**: Hive, Spark SQL, PySpark, EMR, Tableau, Sqoop, AWS, Python, Snowflake, Teradata, Azure AAS & SSAS, Kafka.

**Tower Hill Insurance|** **Gainesville FL January 2019 - August 2020  
Sr Data Engineer**

**Responsibilities:**

* Installing, configuring and maintaining Data Pipelines.
* Transforming business problems into Big Data solutions and defining Big Data strategy and Roadmap.
* Designing the business requirement collection approach based on the project scope and SDLC methodology.
* Authoring Python (PySpark) Scripts for custom UDF's for Row/ Column manipulations, merges, aggregations, stacking, data labeling, and for all Cleaning and conforming tasks.
* Writing Pig Scripts to generate MapReduce jobs and perform ETL procedures on the data in HDFS.
* Develop solutions to leverage ETL tools and identify opportunities for process improvements using Informatica and Python.
* Make recommendations for continuous improvement of the data processing environment.
* Develop a data platform from scratch and took part in the requirement gathering and analysis phase of the project in documenting the business requirements.
* Design and implement multiple ETL solutions with various data sources by extensive SQL Scripting, ETL tools, Python, Shell Scripting, and scheduling tools.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats.
* Used SSIS to build automated multi-dimensional cubes.
* Used Spark Streaming to receive real-time data from Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Validated the test data in DB2 tables on Mainframes and Teradata using SQL queries.
* Automated and scheduled recurring reporting processes using UNIX shell scripting and Teradata utilities such as MLOAD, BTEQ, and Fast Load.
* Worked on Dimensional and Relational Data Modelling using Star and Snowflake Schemas, OLTP/OLAP system, Conceptual, Logical, and Physical data modeling.
* Automated the data processing with Oozie to automate data loading into the Hadoop Distributed File System.
* Developed Automation Regressing Scripts for validation of ETL process between multiple databases like AWS Redshift, Oracle, MongoDB, T-SQL, and SQL Server using Python.

**Environment**: Cloudera Manager (CDH5), Hadoop, Pyspark, HDFS, NiFi, Pig, Hive, S3, Kafka, Scrum, Git, Sqoop, Oozie, Pyspark, Informatica, Tableau, OLTP, OLAP, HBase, Cassandra, Informatica, SQL Server, Python, Shell Scripting, XML, Unix.

**Cummins | Columbus, Indiana April 2016 - December 2018  
Sr Data Engineer**

**Responsibilities:**

* Migrating an entire oracle database to BigQuery and using of power bi for reporting.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Experience in building power bi reports on Azure Analysis services for better performance.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
* Coordinated with team and Developed framework to generate Daily adhoc reports and Extracts from enterprise data from BigQuery.
* Designed and Co-ordinated with Data Science team in implementing Advanced Analytical Models in Hadoop Cluster over large Datasets.
* Wrote scripts in Hive SQL for creating complex tables with high performance metrics like partitioning, clustering and skewing
* Work related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities.
* Worked with google data catalog and other google cloud API’s for monitoring, query and billing related analysis for BigQuery usage.
* Worked on creating POC for utilizing the ML models and Cloud ML for table Quality Analysis for the batch process.
* Knowledge about cloud dataflow and Apache beam.
* Good knowledge in using cloud shell for various tasks and deploying services.
* Created BigQuery authorized views for row level security or exposing the data to other teams.
* Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, SQOOP, Apache Spark, with Cloudera Distribution.

**Environment**: Hive, Spark SQL, PySpark, EMR, Tableau, SQOOP, GCP, Presto, Python, Snowflake, Teradata, Azure AAS & SSAS, Kafka.

**Grape soft Solutions | Hyderabad, India November 2014 - December 2015  
Data Engineer**

**Responsibilities:**

* Extensive knowledge/hands-on experience in architecting or designing Data warehouse/Database, Modelling, building SQL objects such as tables, views, user-defined/ table-valued functions, stored procedures, triggers, and indexes
* Created HBase tables from Hive and Wrote HiveQL statements to access HBase table's data
* Developed complex Hive Scripts for processing the data and created dynamic partitions and bucketing in the hive to improve the query performance
* Developed MapReduce applications using Hadoop Map-Reduce programming framework for processing and used compression techniques to optimize MapReduce Jobs
* Developed Pig UDF's to know the customer behavior and Pig Latin scripts for processing the data in Hadoop
* Scheduled automated tasks with Oozie for loading data into HDFS through Sqoop and pre-processing the data with Pig and Hive
* Develop the Oozie actions like hive, shell, and java to submit and schedule applications to run in the Hadoop cluster
* Worked with the production support team to provide the necessary support for issues with the CDH cluster and the data ingestion
* Worked in Azure environment for development and deployment of Custom Hadoop Applications
* Designed and implemented scalable Cloud Data and Analytical a solution for various public and private cloud platforms using Azure
* Designed and implemented distributed data processing pipelines using Apache Spark, Hive, Python, Airflow DAGs, and other tools and languages in Hadoop Ecosystem.
* **Environment:** Python, MySQL, PostgreSQL, Hadoop (Hive), AWS (S3, EMR), Tableau, Docker, Kafka.

**IBing Software Solutions Private Limited | Hyderabad, India June 2013 - October 2014  
Data Engineer**

**Responsibilities:**

* Built end-to-end ETL models to sort vast amounts of customer feedback, derive actionable insights and tangible business solution
* Developed a data quality control model to monitor business information change overtime
* The model flags outdated customer information using different APIs for validation and updates it with correct data
* Responsible for monitoring sentimental prediction model for customer reviews and ensuring high-performance ETL process
* Data cleaning, pre-processing, and modeling using Python and R
* Develop the automation scripts to transfer the data from on-premise clusters to AWS
* Involved in performance tuning and optimization of long-running MapReduce jobs and queries (Hive/SQL)
* Developed using object-oriented methodology a dashboard to monitor all network access points and network performance metrics using Django, Python, MongoDB, JSON

**Environment:** Python, SQL, AWS (Redshift, S3, EC2), Mongo dB, MYSQL

**EDUCATION:** B.Tech, Electronics and Communication 2013