**RAHUL P**

**Sr. Data Engineer** [**p.rahulp1218@gmail.com**](mailto:p.rahulp1218@gmail.com) **/ PH: 9132133256 LinkedIn:** [www.linkedin.com/in/rahul-p-4b4724247](http://www.linkedin.com/in/rahul-p-4b4724247)

**Professional Summary**

* Data engineering professional having around 8 years of experience in a variety of data platforms, with hands on experience in Big Data Engineering and Data Analytics.
* 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, CloudFormation, CloudTrail, Ops Works, Kinesis, IAM, SQS, SNS, SES.
* Experience in layers of Hadoop Framework - Storage (HDFS), Analysis (Pig and Hive), Engineering (Jobs and Workflows), extending the functionality by writing custom UDFs.
* Define virtual warehouse sizing for Snowflake for different type of workloads.
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Hands-on experience with Snowflake utilities, SnowSQL, SnowPipe, Big Data model techniques using Python / Java.
* Detail-oriented and result driven BI Professional with around 7 years of experience in populating and maintaining Enterprise Data Warehouse and subject area specific Data Marts using IBM DataStage ETL tool.
* Experienced in working with Spark eco system using pyspark and HIVE Queries on different data formats like Text file and parquet.
* 4+ years of experience in BIG DATA Hadoop Administration and Development.
* Used recently introduced Power BI to create self-service Bl capabilities and use tabular models.
* Experience in developing Reports, Charts, Dashboards using Power BI Desktop and manage security based on requirements
* Solid experience in building interactive reports, dashboards, and/or integrating modeling results and have strong Data Visualization design skills and analytic background by using POWER BI and Tableau desktop.
* Extensive knowledge of data architecture including designing pipelines, data ingestion, Hadoop/Spark architecture and advanced data processing.
* Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.
* Reviewed the HDFS usage and system design for future scalability and fault-tolerance. Installed and configured Hadoop HDFS, MapReduce, Pig, Hive, and Sqoop.
* Excellent command in creating Backups & Recovery and Disaster recovery procedures and Implementing BACKUP and RECOVERY strategies for off-line and on-line Backups.
* Worked on data processing and transformations and actions in spark by using Python (AWS) language.
* Experience in collection the real time streaming data and creating the pipeline for data from different source using Kafka and store data into HDFS and NoSQL using Spark.
* Experience in creating SSIS Packages using Pivot Transformation, Fuzzy Lookup, Derived Columns, Condition Split, Term extraction, Aggregate, Execute SQL Task, Data Flow Task, and Execute Package Task etc. to generate underlying data for the reports and to export cleansed data. Experience working with continuous integration framework, building regression-able code within data world using GitHub, Jenkins, and related applications.
* Good experience in managing Kubernetes environment for scalability, availability and zero downtime.
* Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
* 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/Data bricks, and NoSQL DB).
* Design, develop, and test dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
* Experience in working with NoSQL databases like HBase and Cassandra.

**Technical Skills:**

|  |  |
| --- | --- |
| **Operating Systems** | Unix, Linux, Windows |
| **Programming Languages** | Java, Python 3, Scala, Spark, SQL, PL/SQL, UNIX, Pig, HiveQL, Shell Scripting |
| **Hadoop Eco System** | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Apache Flume, Apache Storm, Apache Airflow, HBase, Kafka |
| **Cluster Management & monitoring** | CDH, Horton Works Ambari |
| **Data Bases** | * Oracle 12c/11g/ 10g, MySql, MS Sql, DB2, Snowflake, Redshift, Big Query |
| **NoSQL Data Bases** | MongoDB, Cassandra, HBase, DynamoDB |
| **Workflow mgmt. tools** | Apache Airflow |
| **Visualization & ETL tools** | Tableau, Power BI, Informatica, Talend, SSIS, Ab initio |
| **Cloud Technologies** | Azure, AWS |
| **IDE’s** | Eclipse, Jupyter notebook, Spyder, PyCharm, IntelliJ |
| **Version Control Systems** | Git, SVN, CSV |

**Professional Experience**

**Sr. Data Engineer**

**Travelport Englewood, CO April 2021 to Present**

**Responsibilities:**

* Participated in gathering and identifying data points to create a Data model.
* Worked in Agile methodology in order to discuss status of the project twice a week and take further measures which includes scrum meetings daily to discuss product backlog. Met the expectations by reaching tight deadline on the agile model.
* 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.
* Performed logical and physical data structure designs and DDL generation to facilitate the implementation of database tables and columns out to the DB2, SQL Server, AWS Cloud (Snowflake) and Oracle DB schema environment using Erwin Data Modeler Model Mart Repository version 9.6.
* Served as the Snowflake Database Administrator responsible for leading the data model design and database migration deployment production releases to endure our database objects and corresponding metadata were successfully implemented to the production platform environments; (Dev, Qual, and Prod) AWS Cloud (Snowflake).
* Worked on Informatica Power Center tools- Designer, Repository Manager, Workflow Manager, and Workflow Monitor.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Power BI and SAS Visual Analytics
* Maintained stored definitions, transformation rules and targets definitions using Informatica repository Manager.
* Developed mappings in Informatica to load the data including facts and dimensions from various sources into the Data Warehouse, using different transformations like Source Qualifier, JAVA, Expression, Lookup, Aggregate, Update Strategy and Joiner.
* Install and configure Apache Airflow for S3 bucket and Snowflake data warehouse and created dags to run the Airflow.
* Worked in Snowflake advanced concepts like setting up Resource Monitors, Role Based Access Controls, Data Sharing, Virtual Warehouse Sizing, Query Performance Tuning, Snow Pipe, Tasks, Streams, Zero- copy cloning etc.
* Responsible for creating on-demand tables on S3 files using lambda functions using Python and Pyspark.
* Ingested data into RAW layer using Pyspark framework.
* Created various Parser programs to extract data from Business Objects, XML, Java, and database views using Scala
* Involved in developing python scripts, informatics for ETL tools for extraction, transformation and loading of data into data warehouse.
* Implemented Informatica recommendations, methodologies, and best practices.
* Classified all the data which is coming from source into IVC templates which initiates ingestion based on the classification.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
* Worked on developing Visual reports, Dashboards and KPI scorecards.

**Environment**: AWS EMR/EC2/S3/Redshift/Route 53/SNS, Snowflake, Informatica, Scala, Airflow, Oracle DB, SQL Server, Pyspark, ETL, Power BI, Python, Dynamo DB.

**Sr. Data Engineer**

**GDIT Rensselaer, NY Aug 2019 to March 2021**

**Responsibilities:**

* Processed the Web server logs by developing multi-hop flume agents by using Avro Sink and loaded into MongoDB for further analysis, also extracted files from MongoDB through Flume and processed.
* Expert knowledge on MongoDB, NoSQL data modeling, tuning, and disaster recovery backup used it for distributed storage and processing using CRUD.
* Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in GCP Dataproc, GCS, Cloud functions, Big Query.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Experience in setting up Fan-out workflow in flume to design v shaped architecture to take data from many sources and ingest into single sink.
* Experience in creating tables, dropping, and altered at run time without blocking updates and queries using HBase and Hive.
* Experience in Designing. Architecting and implementing scalable cloud-based web applications using AWS and GCP.
* Set up a GCP Firewall rules in order to allow or deny traffic to and from the VM's instances based on specified configuration and used GCP cloud CD (content delivery network) to deliver content from CP cache locations drastically improving user experience and latency.
* Imported several transactional logs from web servers with Flume to ingest the data into HDFS.
* Worked on continuous Integration tools Jenkins and automated jar files at end of day.
* Worked with Tableau and Integrated Hive, Tableau Desktop reports and published to Tableau Server.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Developed spark applications in python (PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Analyzed the SQL scripts and designed the solution to implement using Scala.
* Used Spark-SQL to Load JSON data and create Schema RDD and loaded it into Hive Tables and handled structured data using Spark SQL.
* Implemented Spark Scripts using Scala, Spark SQL to access hive tables into Spark for faster processing of data.
* 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.
* Experience in loading XML & JSON files into NoSQL databases such as Marklogic and MongoDB Database using Apache Nifi 1.8/1.3.
* Extensively used out of the box Kafka processors available in Nifi to consume data from Apache Kafka specifically built against the Kafka consumer API.
* Generated workflows through Apache Airflow, then Apache Oozie for scheduling the Hadoop jobs which controls large data transformations.
* Design dimensional model, data lake architecture, data vault 2.0 on Snowflake and used Snowflake logical data warehouse for compute.
* Creating Reports in Looker based on Snowflake Connections
* Develop stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
* Tested Apache Tez for building high performance batch and interactive data processing applications on Pig and Hive jobs.
* Setup data pipeline using in TDCH, Talend, Sqoop and PySpark on the basis on size of data loads.
* Designed columnar families in Cassandra and Ingested data from RDBMS, performed transformations and exported the data to Cassandra.

**Environment**: Hadoop (HDFS, Map Reduce), Databricks, Spark, Talend, Impala, Hive, PostgreSQL, Jenkins, Nifi, Scala, Mongo DB, Cassandra, Python, Pig, Sqoop, Hibernate, spring, SailPoint (IBM), Oozie, Snowflake, Gcp, TEZ, MySQL, Oracle.

**Sr. AWS Data Engineer**

**Chewy Dania Beach, FL Dec 2016 to July 2019**

**Responsibilities:**

* Involved in complete SDLC life cycle - Designing, Coding, Testing, Debugging and Production Support.
* Built S3 buckets and managed policies for S3 buckets and used S3 glacier for storage and backup on AWS
* 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.
* Transform and analyze the data using Pyspark, HIVE, based on ETL mappings.
* Develop framework for converting existing PowerCenter mappings and to Pyspark (Python and Spark) Jobs.
* Create Pyspark frame to bring data from DB2 to Amazon S3.
* Performed review and analysis of the detailed system specifications related to the DataStage ETL and related applications to ensure they appropriately address the business requirements.
* Using Data Stage Designer analyzed the source data to Extract & Transform from various source systems (oracle 10g, DB2, SQL server and flat files) by incorporating business rules using different objects and functions that the tool supports.
* Provide guidance to development team working on Pyspark as ETL platform.
* Involved in loading data from UNIX file system to HDFS using Sqoop.
* Worked on setting up pig, Hive and Hbase on multiple nodes and developed using Pig, Hive and HBase, MapReduce.
* Evaluated impact of proposed changes on existing DataStage ETL applications, processes and
* configurations
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions.
* Optimize the Pyspark jobs to run on Kubernetes Cluster for faster data processing.
* Managed Servers on the Amazon Web Services (AWS) platform instances using Puppet configuration management.
* Involved in maintaining the reliability, availability, and performance of Amazon Elastic Compute Cloud (Amazon EC2) instances.
* Using Da ta Stage created mappings and mapplets to transform the data according to the business rules.
* Used Kubernetes to deploy scale, load balance, scale and manage docker containers.
* Secured Hadoop Cluster by implementing Kerberos with Active Directory.
* Used JSON schema to define table and column mapping from S3 data to Redshift
* Created and Configured Workflows and Sessions to transport the data to target warehouse Oracle tables using Data Stage.
* Developed Proof of Concept POC for DataStage to SSIS migration
* Extensively used Control-M scheduler to schedule DataStage jobs
* Import data from different data sources to create Power BI reports and Dashboards.
* Designed, developed, and tested various Power BI visualizations for dashboard and ad-hoc reporting solutions by connecting from different data sources and databases.
* Conduct systems design, feasibility and cost studies and recommend cost-effective cloud solutions such as Amazon Web Services (AWS).
* 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, Oracle, Python, Data Stage, Hadoop, Airflow, Power Bi, Web logic, Unix Shell Scripting, SQL, Kubernetes, docker, Git.

**Data Engineer**

**Hudda InfoTech Private Limited Hyderabad, India April 2015 to Sep 2016**

**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.
* Performed Spark jobs optimization and performance tuning to improve running time and resources.
* Worked on reading and writing multiple data formats like JSON, AVRO, Parquet and ORC on HDFS using Pyspark.
* 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.
* Experience in proving application support for Jenkins.
* Developed a data pipeline with AWS to extract the data from weblogs and store in HDFS.
* Used Hive QL to analyze the partitioned and bucketed data and compute various metrics for reporting.
* Used reporting tools like Tableau to connect with Hive for generating daily reports of data.

**Environments**: Python, Bigdata, Hadoop, HBase, Hive, Spark, Pyspark, Cloudera, Kafka, Airflow, Sqoop, Jenkins, Unix Shell scripting, GitHub, SQL, Tableau.

**Data Analyst**

**Brio Technologies Private Limited Hyd India July 2014 to March 2015**

**Responsibilities:**

* Collaborated with multiple business units to gather requirements and develop data reporting infrastructure to provide real-time business insights.
* Developed Power Bi Data Model using datasets from multiple data sources (Postgres, Oracle and FAST) by establishing meaningful relationships between them, improving performance by 60%.
* Presented data insights to executive stakeholders using dashboards to provide them with data-based recommendations.
* 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.
* Collaborated with QA analysts to troubleshoot data quality issues and worked with data source owners to resolve them; Monitoring dashboard helped improve data quality by 80%
* Utilized Excel features - Pivot Tables, Vlookup to reconcile account values for Actuarial department with Policy Management System reducing risk of invalid cash disbursements by 85%
* Used REST API endpoints to interact with Power Bi web services to automate report deployment across different environments, increasing efficiency by 40%.

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