**Amar**

**Data Engineer**

**PROFESSIONAL SUMMARY:**

* Around 10 years of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a Sr. Data Engineer/Data Developer and Data Modeler.
* Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using Python API, PySpark API and Spark API for analyzing the data.
* Extensively used Python Libraries PySpark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy and Beautiful Soup.
* 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 EMR.
* Hands On experience on Spark Core, Spark SQL, Spark Streaming and creating the Data Frames handle in SPARK with Scala.
* Experience in NoSQL databases and worked on table row key design and to load and retrieve data for real time data processing and performance improvements based on data access patterns.
* Extensive experience in Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and Map Reduce concepts.
* Experience in building large scale highly available Web Applications. Working knowledge of web services and other integration patterns.
* Developed Simple to complex Map/reduce and Streaming jobs using Java and Scala language.
* Developed Hive scripts for end user / analyst requirements to perform ad hoc analysis.
* EMR with Hive to handle less important bulk ETL jobs.
* Hands-on use of Spark and Scala APIs to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement.
* Hands on Spark MLlib utilities such as including classification, regression, clustering, collaborative filtering, dimensionality reduction.
* 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's and Spark YARN.
* Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
* Able to use Sqoop to migrate data between RDBMS, NoSQL databases and HDFS.
* Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
* Worked with various text analytics libraries like Word2Vec, GloVe, LDA and experienced with Hyper Parameter Tuning techniques like Grid Search, Random Search, model performance tuning using Ensembles and Deep Learning.
* Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS specific features.
* Knowledge of working with Proof of Concepts (PoC's) and gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging and Teradata.
* Well experience in Normalization and De-Normalization techniques for optimum performance in relational and dimensional database environments.
* Experience in developing customized UDF’s in Python to extend Hive and Pig Latin functionality.
* Expertise working with AWS cloud services like EMR, S3, Redshift, EMR cloud watch, for big data development.
* Good 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.
* Utilized **AWS CLI** to automate backups of ephemeral data-stores to **S3 buckets**, **EBS** and create nightly AMIs for mission critical production servers as backups.
* Experience in migrating infrastructure and application from onpremise to Azure and from Cloud to Cloud such as AWS to Microsoft Azure.
* Experience working on Azure Cloud services, Azure Data Factory, Azure Data Lake Analytics, Azure Data Bricks, GIT, Azure DevOps. Managing Client’s Microsoft Azure-based PAAS and IAAS environment.
* Extensive experience in WindowsAzure (IaaS) migrating like creating Azure VMs, storage accounts, VHDs, storage pools, migrating on premise servers to Azure and creating availability sets in Azure.
* Experience in dealing with Windows Azure IaaS - Virtual Networks, Subnets, Load balancers, Virtual Machines, Cloud Services, Resource Groups, Express Route, VPN, Load Balancing, Application Gateways, Auto-Scaling, and Traffic Manager.
* Designed and implemented real-time data pipelines incorporating GenAI components, enabling the transformation of data into actionable insights instantaneously.
* Experience in Data Analysis, Data Profiling, Data Integration, Migration, Data governance and Metadata Management, Master Data Management and Configuration Management.
* Experience in developing customized UDF’s in Python to extend Hive and Pig Latin functionality.
* Expertise in designing complex Mappings and have expertise in performance tuning and slowly changing Dimension Tables and Fact tables
* Experienced in designing roles and groups for users and resources using AWS Identity Access Management (IAM). Managed multiple AWS accounts with multiple VPC's for both production and nonproduction where primary objectives included automation, build out, integration and cost control.
* Extensively worked with Teradata utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files.
* Experienced in designing roles and groups for users and resources using AWS Identity Access Management (IAM). Managed multiple AWS accounts with multiple VPC's for both production and nonproduction where primary objectives included automation, build out, integration and cost control.
* Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
* Leveraged GenAI solutions to build predictive analytics models that analyze real-time data streams, providing valuable insights for optimizing business processes.
* Expert in building Enterprise Data Warehouse or Data warehouse appliances from Scratch using both Kimball and Inmon’s Approach.
* Experience in designing star schema, Snowflake schema for Data Warehouse, ODS architecture.
* Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS specific features.
* Well experience in Normalization and De-Normalization techniques for optimum performance in relational and dimensional database environments.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Big data Technologies | HDFS, Map Reduce, Pig, Hive, Sqoop, Oozie, Scala, Spark, Kafka, Airflow, Flume, Ambari, Hue |
| Hadoop Frameworks | Cloudera CDHs, Hortonworks HDPs, MAPR |
| Database | Oracle 10g/11g, PL/SQL, MySQL, MS SQL Server 2012, DB2 |
| Language | C, C++ , Scala, Python |
| Cloud & DevOps | Continuous Integration & Delivery: Atlassian Bamboo, GitHub Actions |
| AWS Components | IAH, S3, EMR, EC2, Lambda, Redshift, |
| Methodologies | Agile, Waterfall |
| Build Tools | Maven, Gradle, Jenkins |
| Databases | NO-SQL, HBase, Cassandra, MongoDB, DynamoDB |
| IDE Tools | Eclipse, Net Beans, IntelliJ |
| Modelling Tools | Rational Rose, Star UML, Visual paradigm for UML |
| BI Tools | Tableau |
| Operating System | Windows 7/8/10, Vista, UNIX, Linux, Ubuntu, Mac OS X |

**Premera BCBS, Mountlake Terrace, Washington, United States**

**Role: Data Engineer May 2020 - Present**

**Responsibilities:**

* Performed data analysis and developed analytic solutions. Data investigation to discover correlations / trends and the ability to explain them.
* Worked with Data Engineers, Data Architects, to define back-end requirements for data products (aggregations, materialized views, tables – visualization)
* Developed frameworks and processes to analyze unstructured information. Assisted in Azure Power BI architecture design
* Experienced with machine learning algorithm such as logistic regression, random forest, XGboost, KNN, SVM, neural network, linear regression, lasso regression and k - means
* Implemented Statistical model and Deep Learning Model (Logistic Regression, XGboost, Random Forest, SVM, RNN, CNN).
* The purpose of Grafana dashboards is to bring data together in a way that is both efficient and organized.
* Designing and Developing Oracle PL/SQL and Shell Scripts, Data Import/Export, Data Conversions and Data Cleansing.
* It allows users to better understand the metrics of their data through queries, informative visualizations and alerts.
* Responsible for importing data from PostgreSQL to HDFS, HIVE using SQOOP tool.
* Experienced in migrating HiveQL into Impala to minimize query response time.
* Implemented Avro and parquet data formats for apache Hive computations to handle custom business requirements.
* Design & development of DataStage jobs using IBM InfoSphere DataStage Designer to generate pipe-delimited extract files that is FTP’ed to vendor of New York State
* Wrote complex Functions in GreenPlum (database) querying FACETS data including performance tuning that is invoked by DataStage jobs.
* involved in designing and deploying multi-tier applications using all the AWS services.
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Created views in AWS Athena to allow secure and streamlined data analysis access to downstream business teams.
* Created AWS cloud formation templates to create organization AWS network from scratch with customized VPC, subnets, NAT gateways, internet gateways, Route tables, ACL’s EC2 instances, ELB's, security groups.
* Managed multiple AWS accounts with multiple VPC's for both production and non-prod where primary objectives included automation, build out, integration and cost control.
* Designed roles and groups for users and resources using AWS Identity Access Management (IAM) and managed IAM account (with MFA) and IAM policies to meet security audit & compliance requirements and managed network security using Security Groups and IAM.
* Writing AWS Terraform templates for any automation requirements in AWS services. Implemented multiple high-performance MongoDB replica sets on EC2 with robust reliability. Managed and supported AWS Security related issues, such IAM and S3 policies for user access.
* Developed and maintained highly scalable and fault-tolerant multi-tier AWS environments spanning across multiple availability zones using Terraform.
* Involved in various aspects and phases of architecting, designing, and implementing solutions in IT infrastructure with an emphasis on Azure and AWS cloud.
* Developing Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats. Using Kafka and integrating with the Spark Streaming. Developed data analysis tools using SQL and Python code.
* Automate Deployment using Configuration Management tool like puppet to provision amazon AWS Instances to enable continuous Deployments.
* Worked on Cloud automation using AWS Cloud Formation templates.
* Deploy and monitor scalable infrastructure on Amazon web services (AWS) & configuration management using puppet.
* For organizations with critical machinery or equipment, GenAI can be used to implement predictive maintenance. Real-time data from sensors on machines is analyzed to predict when maintenance is required, reducing downtime and maintenance costs.
* Involved in creating external Hive tables from the files stored in the AWS S3.
* 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.
* Grafana Enterprise Metrics provides a centralized, horizontally scalable, replicated architecture.
* Authoring Python (PySpark) Scripts for custom UDF’s for Row/ Column manipulations, merges, aggregations, stacking, data labeling and for all Cleaning and conforming tasks. Migrate data from on-premises to AWS storage buckets.
* Implemented change data capture (CDC) using DataStage to enable near real-time data updates, facilitating quicker decision-making and reducing data staleness in downstream applications.
* Scheduled, deployed, and managed container replicas onto a node cluster using Kubernetes.
* so you can easily manage and maintain your Prometheus implementation based on your specific architecture.
* Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designer.
* Deployed the Big Data Hadoop application using Talend on cloud AWS (Amazon Web Services) and on Microsoft Azure.
* Worked on Azure PaaS Solutions like Azure Web Apps, Web Roles, Worker Roles, SQLAzure and Azure Storage. And managing Azure Storage Accounts and Creating ARM templates for Azure Platform.
* Worked on Migrating servers, databases, Docker, and applications from on premise to Azure.
* Working on Microsoft Azure in creating cloud environment and managing the application servers across Microsoft managed global data cent.
* Designed and configured AZURE Virtual Networks, subnets, network settings, DHCP address blocks, DNS settings, and Security policies and routing.
* Configured storage account, a vault, and a virtual network in Azure, enabled replication for individual VM, and monitored automated mechanisms within Azure site recovery (ASR) during the migration of AWS based services onto Azure.
* Setup Azure Devops for VMs applications migrating from on-prem. Azure Boards and Enterprise Git integration. YAML files for automation and CI/CD.ARM templates for managing the VMs, Key Vaults, Storage accounts.
* Implemented high availability with Azure Classic and Azure Resource Manager deployment models.
* Developed Mappings using Transformations like Expression, Filter, Joiner and Lookups for better data messaging and to migrate clean and consistent data
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* 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.
* Responsible for loading processed data to AWS Redshift table for allowing Business reporting team to build dashboards.
* Develop scripts to deploy jobs using oracle ​DBMS\_SCHEDULAR​to gather statistics on regular intervals to get better Execution plans
* 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
* 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.
* Used Oracle 10g new feature data pump utilities (EXPDP/IMPDP) for Copying of data from development server to test and production server using.
* Used Apache NiFi to copy data from local file system to HDP.
* Worked on Dimensional and Relational Data Modeling using Star and Snowflake Schemas, OLTP/OLAP system, Conceptual, Logical and Physical data modeling using Erwin.
* Automated the data processing with Oozie to automate data loading into the Hadoop Distributed File System
* Built the machine learning model include: SVM, random forest, XGboost to score and identify the potential new business case with Python Scikit-learn.
* Experience in Converting existing AWS Infrastructure to Server less architecture (AWS Lambda, Kinesis), deploying via Terraform and AWS Cloud Formation templates.
* Worked on Docker containers snapshots, attaching to a running container, removing images, managing Directory structures and managing containers.
* Wrote PL/SQL procedures, packages, triggers in Oracle 9i, Involved in creation of databases, moved databases by recreating control files, export/import, and complete backups. Created and maintained Oracle schema objects like Table spaces, Tables, Indexes, Sequences and Synonyms.
* Used Jenkins pipelines to drive all microservices builds out to the Docker registry and then deployed to Kubernetes, Created Pods and managed using Kubernetes
* Experienced in day - to-day DBA activities including schema management, user management (creating users, synonyms, privileges, roles, quotas, tables, indexes, sequence), space management (table space, rollback segment), monitoring (alert log, memory, disk I/O, CPU, database connectivity), scheduling jobs, UNIX Shell Scripting.
* Expertise in using Docker 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. Developed Talend ESB services and deployed them on ESB servers on different instances.
* Architect and design serverless application CI/CD by using AWS Serverless (Lambda) application model.
* Utilized Kubernetes for the runtime environment of the CI/CD system to build, test deploy.
* Developed stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
* Developed merge scripts to UPSERT data into Snowflake from an ETL source.
* **Environment:** Hadoop, Map Reduce, HDFS, Hive, Sqoop, Spring Boot, Cassandra, Swamp, Data Lake, Sqoop, Oozie, Kafka, Spark, Scala, Java, AWS, GitHub, Docker, Talend Big Data Integration, Solr, Impala, Oracle, SQL Server, MySQL, No SQL, MongoDB, HBase, Cassandra, UNIX, Shell Scripting,

**AbbVie Chicago, Illinois**

**Role : Data Engineer Sep 2019 - May 2020**

**Responsibilities:**

* Developed methods for ingesting data from different sources and processing Data-at-Rest utilizing Big Data tools like Hadoop, Map Reduce Frameworks, HBase, and Hive.
* Using best practices, Snowflake was deployed, and subject matter experts in data warehousing, especially with Snowflake, were provided.
* A workflow engine called Oozie was created to handle numerous Hive, Pig, Tealeaf, Mongo DB, Git, Sqoop, and Spark processes.
* Utilizing Informatica Designed, a variety of Mappings were created using the collection of all Sources, Targets, and Transformations.
* POC using Tableau and AWS Quick sight are conducted to suggest product direction. created a dashboard and a metadata layer.
* Use HDFS and Sqoop command to handle extensive data input. gather "partitioned" data in a variety of storage formats, such as text, JSON, Parquet, etc. engaged in the loading of data from the LINUX file system to the HDFS.
* Configuring and integrating the required AWS services in accordance with the business requirement to start Infrastructure as a code (Iaas) in the AWS cloud platform from scratch.
* Created and crafted the study enrollment view for the metadata layer to use.
* Data analysis and visualization were done using AWS Athena and quick sight.
* Participating in all phases of the project and its scope, reference data for MDM was used to develop a Data Dictionary and a Mapping from Sources to the Target in the MDM Data Model.
* Creating a data pipeline using Spark, Scala, and Apache Kafka to ingest data from a CSL source and store it in an HDFS secured folder.
* Designing and deploying multi-tier applications with an emphasis on high availability, fault tolerance, and auto-scaling on AWS Cloud Formation utilizing all of the AWS services (EC2, AWS GLUE, Athena, Lambda, S3, RDS, Dynamo DB, SNS, SQS, IAM, etc.).
* Designing and deploying several applications that make use of practically all AWS services, with an emphasis on high availability, fault tolerance, and auto-scaling in AWS Cloud Formation, including EC2, RedShift, S3, RDS, Dynamo DB, SNS, and SQS.
* Custom Kafka producers and consumers have been developed for a variety of publishing and subscribing to Kafka topics.
* Using sophisticated SQL scripting, ETL tools, Python, Shell scripting, and scheduling tools, create and deploy several ETL systems with different data sources. XML, Web feeds, and file processing utilizing Python, Unix, and SQL for data profiling and manipulation.
* Data migration between many Teradata servers was done to assist the development team.
* Architecting and creating serverless web applications with AWS Lambda, API Gateway, Dynamo DB, and Security Token Service (STS).
* Linked to several stages of the Software Development Life Cycle (SDLC) of the program, including requirement collecting, design, analysis, and code development.
* Create data models for the apps, views, tables, and other database objects that contain metadata.
* Build automated ETL pipelines that can process a wide range of data sources.
* Knowledgeable about loading and extracting data from Python sources and working with Python tools for data analysis, including matplotlib, NumPy, SciPy, and Pandas
* Building Python-based database models, APIs, and views to create an interactive web-based solution.
* Coordinating the development of the team using the GIT version control tool.
* Jenkins and SonarQube were integrated to enable SonarQube's Maven scanner to do continuous code quality inspection and analysis.
* Using Spark RDDs and Scala to transform Hive/SQL queries into Spark Transformations, as well as SQOOP to import and export data between RDBMS and HDFS
* Created dimensional data models utilizing star and snowflake schemas as well as 3NF data models for OLTP systems.
* Utilizing Terraform scripts from Jenkins, involved in provisioning AWS infrastructure.
* Docker and Kubernetes were heavily utilized to securely ship, run, and deploy the application in containers to speed up the build and release engineering process.
* Sqoop tasks were made to import the data from DB2 to HDFS.
* Created visuals, processed XML, exchanged data and implemented business logic using Python and Django.

**Zurich insurance, Schaumburg, IL**

**Data Engineer Sep 2018 – Aug 2019**

**Roles & Responsibilities:**

* Creating framework for the data Ingestion from various sources to Hadoop using python and spark
* Creating Test Automation Framework using Scala
* Creating utility method to flatten json events to granular level using Scala & Spark
* Performed end-to-end delivery of pyspark ETL pipelines on Azure-databricks to perform the transformation of data orchestrated via Azure Data Factory (ADF) scheduled through Azure automation accounts and trigger them using Tidal Schedular.
* Data modeled HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of data sources.
* Solved performance issues in Hive scripts with understanding of Joins, Group and aggregation and translate to MapReduce jobs.
* Developed UDFs in Java as and when necessary to use in HIVE queries.
* Coordinated with various stakeholders such as the End Client, DBA Teams, Testing Team and Business Analysts.
* Involved in gathering requirements and developing a project plan.
* Involved in understanding requirements, functional specifications, designing documentations and testing strategies.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD's.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Created Azure functions and configured it to receive events from your Synapse warehouse.
* Implemented data streaming capability using Kafka and Talend for multiple data sources.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Migrated an existing on-premises application to Azure.
* Used spark sql to load data and created schema RDD on top of that which loads into hive tables and handled structured using spark Sql.
* Involved in UI designing, Coding, Database Handling.
* Used Azure services like ADLS and Synapse analytics for small data sets.
* Involved in Unit Testing and Bug Fixing.
* Created Azure functions and configured it to receive events from your Synapse warehouse.
* Worked over the entire Software Development Life Cycle (SDLC) as a part of a team as well as independently.
* Written SQL queries to query the database and providing data extracts to users as per request

**Environment:**

* SQL server, Teradata, Mongo DB, Oracle Putty, WINSCP, SFTP, Hadoop (Cloudera) cluster, mapR cluster, Jupiter notebook, PyCharm, IntelliJ, bitbucket, bamboo, AzurevAvro, json, XML, Unix, Python, Scala, PL/SQL, Ansible, spark, hdfs, hive, HBase, Sqoop, kafka, spark streaming, spark-sql, Jenkins, AWS/Azure models

**peloton interactive inc**

**Role :Data Engineer June 2016 -Aug 2018**

**Responsibilities:**

* As a Data Engineer I am responsible for building scalable distributed data solutions using Hadoop.
* Involved in Agile Development process (Scrum and Sprint planning).
* Handled Hadoop cluster installations in Windows environment.
* Migrated on-premise environment in GCP (Google Cloud Platform)
* Migrated data warehouses to Snowflake Data warehouse.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Involved in porting the existing on-premise Hive code migration to GCP (Google Cloud Platform) BigQuery.
* Involved in migration an Oracle SQL ETL to run on Google cloud platform using cloud Dataproc & BigQuery, cloud pub/sub for triggering the Apache Airflow jobs.
* Extracted data from data lakes, EDW to relational databases for analyzing and getting more meaningful insights using SQL Queries and PySpark.
* Developed PySpark script to merge static and dynamic files and cleanse the data.
* Created Pyspark procedures, functions, packages to load data.
* Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
* Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
* Set up Data Lake in Google cloud using Google cloud storage, BigQuery and Big Table.
* Developed scripts in BigQuery and connecting it to reporting tools.
* Designed workflows using Airflow to automate the services developed for Change data capture.
* Carried out data transformation and cleansing using SQL queries and PySpark.
* Used Kafka and Spark streaming to ingest real time or near real time data in HDFS.
* Worked related to downloading BigQuery data into Spark data frames for advanced ETL capabilities.
* Worked on PySpark APIs for data transformations.
* Built reports for monitoring data loads into GCP and drive reliability at the site level.
* Participated in daily stand-ups, bi-weekly scrums and PI panning.

**Environment**: Hadoop, Map Reduce, HDFS, Hive, Sqoop, Spring Boot, Cassandra, Swamp, Data Lake, Sqoop, Oozie, Kafka, Spark, Scala, Java, AWS, GitHub, Docker, Talend Big Data Integration, Solr, Impala, Oracle, SQL Server, MySQL, No SQL, MongoDB, HBase, Cassandra, UNIX, Shell Scripting,

**Volkswagen Group Sales India Pvt Ltd**

**Role: Data Analyst Feb 2014 – May 2016**

**Responsibilities:**

* Imported Legacy data from SQL Server and Teradata into Amazon S3.
* Created consumption views on top of metrics to reduce the running time for complex queries.
* Exported Data into Snowflake by creating Staging Tables to load Data of different files from Amazon S3.
* As a part of Data Migration, wrote many SQL Scripts for Mismatch of data and worked on loading the history data from Teradata SQL to snowflake.
* Developed SQL scripts to Upload, Retrieve, Manipulate and handle sensitive data (National Provider Identifier Data I.e., Name, Address, SSN, Phone No) in Teradata, SQL Server Management Studio and Snowflake Databases for the Project
* Worked on to retrieve the data from FS to S3 using spark commands
* Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage and backup on AWS
* Created performance dashboards in Tableau/ Excel / Power point for the key stakeholders
* Incorporated predictive modeling (rule engine) to evaluate the Customer/Seller health score using python scripts, performed computations and integrated with the Tableau viz.
* Worked with stakeholders to communicate campaign results, strategy, issues or needs.
* Analyzed marketing campaigns from various perspectives including CTR, conversion rates, seasonal/geographical trends, search queries, landing page, conversion funnel, quality score, competitors, distribution channel, etc. to achieve maximum ROI for clients.
* Understood Business requirements to the core and came up with Test Strategy based on Business rules
* Implemented Defect Tracking process using JIRA tool by assigning bugs to Development Team
* Involved in Functional Testing, Integration testing, Regression Testing, Smoke testing and performance Testing. Tested Hadoop Map Reduce developed in python, pig, Hive
* Created Metric tables, End user views in Snowflake to feed data for Tableau refresh.
* Generated Custom SQL to verify the dependency for the daily, Weekly, Monthly jobs.
* Using Nebula Metadata, registered Business and Technical Datasets for corresponding SQL scripts
* Experienced in working with spark ecosystem using Spark SQL and Scala queries on different formats like text file, CSV file.
* Developed spark code and spark-SQL/streaming for faster testing and processing of data.
* Closely involved in scheduling Daily, Monthly jobs with Precondition/Postcondition based on the requirement.

**Environment**: Snowflake, Hadoop, Map Reduce, Spark SQL, Python, Pig, AWS, GitHub, EMR, Nebula Metadata, Teradata, SQL Server, Apache Spark, Sqoop

**BitSys Infotech, India**

**Role : Software Engineer Oct 2012 – Jan 2014**

**Responsibilities:**

* Created a POC in Big Data from end-to-end using Pig, HIVE, HDFS, Tableau
* End to End development of Hadoop Project
* Prepared PL/SQL packages and procedures for the back end processing of the proposed database design
* Delivered PL/SQL training session for co-workers to educate about the latest PL/SQL features, PL/SQL performance tuning
* Drafted tables, synonyms, sequences, views, PL/SQL stored procedures and triggers
* Facilitated testing and code review
* Performed performance tuning of the overall system by eliminating redundant joins, creating indexes, removing redundant code
* Developed UNIX shell scripts for part processing
* Utilized Oracle Designer 6i to perform data modelling
* Documented Tech Specs for the proposed database design
* Devised PL/SQL packages and procedures for the back-end processing of the proposed database design
* Delivered PL/SQL training session for co-workers to educate about the latest PL/SQL features, PL/SQL performance tuning
* Facilitated management of database
* Designed tables, synonyms, sequences, views, PL/SQL stored procedures and triggers
* Performed testing and code review
* Conducted performance tuning of the overall system by eliminating redundant joins, creating indexes, removing redundant code. Developed UNIX shell scripts to perform a nightly refresh of the test system from Production databases. Monitored user profiles, roles and privileges for the Sybase database
* Maintaining technical, functional documentation and all the deliverables.

**Environment:**

* Oracle
* Forms Developer, Report Developer, HP UX 11i, UNIX SUN Solaris 5.8, Oracle 9i, Putty, WINSCP
* Pl/SQL, SQL\*Plus,Unix scripting