Venkatesh P
Gmail: venkatp.data@gmail.com | +1(313) 409-8399

SUMMARY:

- **9+** years of experience in Analyzing, Designing, Developing and Implementation of data, architecture, frameworks as a Senior Data Engineer.
- Specialized in Data Warehousing, Decision support Systems and extensive experience in implementing Full Life cycle Data Warehousing Projects and in Hadoop/Big Data related technology experience in Storage, Querying, Processing, analysis of data.
- Software development involving cloud computing platforms like Amazon Web Services (AWS), Azure and Google Cloud (GCP).
- Expertise in using SSIS and Informatica to extract, transform, and load data from a variety of sources and targets.
- Created reusable Snowflake UDFs, Stored Procedures and Views to standardize constructs across multiple pipelines.
- Established near-zero maintenance security policies, access controls and MFA for Snowflake account.
- Developed and maintained ETL processes to move data from source systems to Snowflake.
- Strong proficiency in SQL concepts, Presto SQL, Hive SQL, Python (Pandas, NumPy, SciPy, Matplotlib), Scala, Java, and Spark to handle large volumes of data.
- Expertise in Database design and development with Business Intelligence using SQL Server 2014/2016, Integration Services (SSIS), DTS Packages, SQL Server Analysis Services (SSAS), DAX, OLAP Cubes, Star Schema and Snowflake Schema
- Knowledge in installing, configuring, and using Hadoop ecosystem components like Hadoop Map Reduce, HDFS, HBase, Oozie, Hive, Sqoop, Zookeeper and Flume.
- Developed and implemented data transformation strategies using SAP Data Services, LSMW, BDC, and other SAP tools.
- Hands-on experience in designing and implementing data engineering pipelines and analyzing data using AWS stack like AWS EMR, AWS Glue, EC2, AWS Lambda, Athena, Redshift, Sqoop and Hive.
- Participated in development/implementation of Cloudera Hadoop environment.
- Good experience working with various data analytics in AWS Cloud like EMR, Redshift, S3, Athena,
- Hands on experience in using various Hadoop distros (Cloudera (CH 4/CDH 5), Hortonworks, Map-R, IBM Big Insights, Apache and Amazon EMR Hadoop distributions.
- 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.
- Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems like Teradata, Oracle, SQL Server and vice-versa.
- Developed and maintained SSIS packages to extract, transform, and load (ETL) data from a variety of sources, including flat files, relational databases, and XML files.
- Used SSIS to automate data movement and processing, including data cleaning, validation, and aggregation.
- Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, Amazon Elastic Load Balancing, and other services in AWS. Deep understanding of Cloud Architectures including AWS, Azure, GCP.
- Used export and import from Snowflake and wrote complex Snowsql scripts in Snowflake cloud data warehouse to business analysis and reporting.
- Adding columns to snowflake views, creating Snowflake tables by accessing data from Taradata.
- Strong proficiency in SQL concepts, Presto SQL, Hive SQL, Python (Pandas, NumPy, SciPy, Matplotlib), Scala, and Spark to handle large volumes of data.
- Skilled in Shell/Bash scripting and building data pipelines on Unix/Linux systems.

Technical Skills:

Big Data Tools	Hadoop, Hive, Apache Spark, PySpark, HBase, Kafka, YARN, Git Hub,Sqoop, Impala, Oozie, Pig, Map Reduce, Zookeeper and Flume
Hadoop Distributions	EMR, Cloudera, Hortonworks.
Database	Snowflake, Oracle, MySQL, SQL Server, PostgreSQL, Teradata, Spark-Redis, DB2
Cloud Services	AWS – EC2, S3, EMR, RDS, Glue, Presto, Lambda, RedShift and Azure – Data Lakes, AWS (EC2, S3, RDS, ECS, EKS, CloudFormation, IAM, VPC, Route 53), BLOB
GCP Cloud Platform	Big Query, Cloud Data Proc, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Shell, GSUTIL, BQ Command Line, Cloud Data Flow.
BI and Data Visualizations	ETL -Informatica, SSIS, Talend, Tableau and Power BI
Relational Databases	Oracle, SQL Server, Teradata, MySQL, PostgreSQL and Netezza
No SQL Databases	Cassandra, MongoDB and HBase , Amazon EKS
Programming Languages	Scala, Python and R
Scripting	Python and Shell scripting
Build Tools	Apache Maven and SBT,Kubernets, Jenkins, Bitbucket
Version Control	GIT and SVN
Operating Systems	Unix, Linux, Mac OS, CentOS, Ubuntu and Windows
Tools	PUTTY, Putty-Gen, Eclipse, Open shift , IntelliJ and Toad
Methodologies	RAD, JAD, UML, System Development Life Cycle (SDLC), Jira, Agile, Confluence, and Waterfall Model

Professional Experience

Client: CVS Healthcare, TX **Oct2022-Present.**
Role: Sr. Data Engineer

Responsibilities:

- Developed Spark programs to parse the raw data, populate staging tables, and store the refined data in partitioned tables in the Enterprise Data warehouse.
- Developed Streaming applications using PySpark to read from the Kafka and persist the data NoSQL databases such as HBase and Cassandra.
- Implemented PySpark Scripts using Spark SQL to access hive tables into a spark for faster processing of data.
- Performed upgrades, scaling actions and zero-downtime migrations for Snowflake deployments handling over 100k DAU.
- Carried out POCs to validate adoption of Snowflake for analytics, data science and data lake use cases.
- Developed and implemented an Amazon EKS solution for containers, optimizing deployment processes and ensuring scalability.
- Provided expert support for OpenShift clusters, effectively addressing customer inquiries, requests, and issues to maintain high satisfaction levels.
- Demonstrated strong troubleshooting skills by efficiently resolving functional issues in the code, minimizing downtime and enhancing system reliability.
- Proficient in breaking down complex problems into manageable tasks and defining clear acceptance criteria, leading to improved project outcomes.

- Proactively engaged in team and project planning activities, contributing valuable insights and fostering collaboration for successful project execution.
- evangelized best practices for Snowflake among developers and analysts to optimize working with cloud data platform.
- Worked on Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
- Implemented Responsible AWS solutions using EC2, S3, RDS, EBS, Elastic Load Balancer, and Auto scaling groups, Optimized volumes and EC2 instances.
- Wrote Terraform templates for AWS Infrastructure as a code to build staging, production environments & set up build & automation for Jenkins.
- Developed and optimized Snowflake data models for 30+ analytics use cases across sales, marketing and finance departments.
- Collaborated with cross-functional teams to design and implement an innovative Amazon EKS solution, enhancing container deployment efficiency by X%.
- Led troubleshooting efforts to identify and resolve complex functional issues in the code, ensuring system stability and optimal performance.
- Designed and implemented scalable cloud-native solutions on AWS, leveraging services such as Amazon EKS, AWS Lambda, and Amazon S3 to optimize performance and cost efficiency.
- Led the migration of on-premises workloads to AWS cloud-native architecture, resulting in improved scalability, reliability, and security of applications.
- Developed CI/CD pipelines using AWS CodePipeline and AWS CodeBuild to automate build, test, and deployment processes, reducing manual errors and deployment time.
- Implemented serverless architectures on AWS using AWS Lambda functions and API Gateway, leading to significant cost savings and improved resource utilization.
- Collaborated with cross-functional teams to architect cloud-native solutions that align with best practices and industry standards, ensuring high availability and scalability of applications.
- Demonstrated expertise in system architecture, contributing to the seamless integration of core systems within the container platform environment.
- Actively participated in continuous improvement initiatives, driving operational excellence and efficiency within the container platform team.
- Leveraged strong communication skills to liaise with customers, address their inquiries, and provide technical support for OpenShift and Amazon EKS clusters.
- Loaded over 5TB of structured and semi-structured data into Snowflake from S3, Kafka and DBs using Tasks and Streams.
- Migrating an entire oracle database to Big Query and using powerBI for reporting.
- Developed streaming and batch processing applications using PySpark to ingest data from the various sources into HDFS Data Lake.
- Developed DDLs and DMLs scripts in SQL and HQL for analytics applications in RDBMS and Hive.
- Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in the UNIX operating system as well.
- Worked on AWS Cloud to convert all premises, existing processes and databases to AWS Cloud.
- Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external Installed and configured Hadoop MapReduce HDFS Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
- Conducted in-depth analysis of system architecture and requirements to design and implement a robust Amazon EKS solution, improving deployment speed .
- Implemented proactive monitoring tools to detect and address system issues promptly, ensuring high availability and performance of OpenShift and Amazon EKS clusters.
- Collaborated with stakeholders to define acceptance criteria for projects, leading to successful project completion within scheduled timelines.
- Demonstrated autonomy and initiative in resolving technical challenges, showcasing strong problem-solving skills in various environments.
- Actively contributed to knowledge sharing sessions within the team, fostering a culture of continuous learning and development in container platform technologies.

- Worked on various data modeling concepts like star schema, and snowflake schema in the project.
- Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and performance analysis.
- Developed Python scripts to find vulnerabilities with SQL Queries by doing SQL injection.
- Used Oozie Scheduler systems to automate the pipeline workflow and orchestrate the map-reduce jobs that extract and Zookeeper for providing coordinating services to the cluster.
- Created and implemented SSIS error handling and logging mechanisms to ensure the smooth operation of data integration pipelines.

Environment: CDH5, Hortonworks, Apache Hadoop 2.6.0, HDFS, Java 8, Hive 1.2.1000, Sqoop 1.4.6, HBase 1.1.2, Oozie 4.1.0, Storm 0.9.3, YARN, NiFi, Cassandra, Zookeeper, Spark, Kafka, Oracle 11g, MySQL, Shell Script, EC2, snowflake, Python, Tomcat 8, Spring 3.2.3, STS 3.6, Build Tool Gradle 2.2, Source Control GIT, Tera Data SQL Assistant

Client: Fifth Third Bank - Cincinnati, Ohio **Nov 2020-Sep2022**
Role: Sr. Data Engineer

Responsibilities:
- Involved in full Software Development Life Cycle (SDLC) - Business Requirements Analysis, preparation of Technical Design documents, Data Analysis, Logical and Physical database design, Coding, Testing, Implementing, and deploying to business users.
- Involved in full Software Development Life Cycle (SDLC) - Business Requirements Analysis, preparation of Technical Design documents, Data Analysis, Logical and Physical database design, Coding, Testing, Implementing, and deploying to business users.
- Transformed and analyzed the data using Pyspark, HIVE, based on ETL mappings.
- Developed spark programs and created the data frames and worked on transformations.
- Analyzed large and critical datasets using HDFS, HBase, Hive, Scala, HQL, PIG, Sqoop, Kubernetes and Zookeeper.
- Develop dashboards in poker to visualize teh suspicious patterns/activities in real time for business users.
- Spearheaded the migration from OpenShift to Amazon EKS, ensuring a smooth transition and minimal disruption to application teams.
- Conducted performance optimization initiatives for container deployment, resulting in a significant reduction in system latency and improved overall efficiency.
- Implemented robust security measures for OpenShift and Amazon EKS clusters, enhancing data protection and compliance with industry standards.
- Acted as a subject matter expert in container platform technologies, providing guidance and mentorship to team members on best practices and emerging trends.
- Developed and delivered comprehensive training programs on OpenShift and Amazon EKS for internal stakeholders, increasing team proficiency and adoption rates.
- Integrated data from multiple sources into SAP.
- Orchestrated the deployment of microservices architecture on AWS using Amazon ECS, enhancing scalability and fault tolerance of applications.
- Implemented AWS CloudFormation templates to automate infrastructure provisioning and configuration management, streamlining the deployment process.
- performance tuning efforts for AWS cloud-native applications, identifying and resolving bottlenecks to enhance overall system efficiency.
- Collaborated with security teams to implement best practices for securing cloud-native environments on AWS, ensuring data integrity and compliance with regulatory requirements
- Tested data transformations to ensure accuracy and completeness.
- Good at creating a Star/Snowflake scheme depending on the requirement and creating complex views or stored procedure to design the required fact & dimension tables for the tabular model and then establishing relationships.

- Developed data pipeline using Flume, Kafka, and Spark Stream to ingest data from their weblog server and apply the transformation.
- Performed data analysis and profiling of source data to better understand the sources.
- Work related to downloading Big Query data into pandas or Spark data frames for advanced ETL capabilities.
- Carried out data transformation and cleansing using SQL queries, Python and Pyspark.
- Wrote scripts in Hive SQL for creating complex tables with high-performance metrics like partitioning, clustering, and skewing.
- Created ETL Pipeline using Spark and Hive for ingest data from multiple sources.
- Was responsible for ETL and data validation using SQL Server Integration Services.
- Reverse-engineered existing data models to incorporate new changes utilizing Erwin.
- Developed artifacts that are consumed by the data engineering team such as source-to-target mappings, data quality rules, data transformation rules, Joins, etc.
- Develop and deploy the outcome using spark and scala code in Hadoop cluster running on GCP.
- Extensively used spark SQL and Data frames API in building spark applications.

Environment: Informatica Power Center 9.5, Data proc,Looker, Snowflake. Scala, Apache spark, Talend, Google Cloud Platform(GCP), PostgreSQL Server, Python, Oracle, Teradata, CRON, Unix Shel Scripting, SQL, Erwin, GitHub

Client: Molina Healthcare, Bothell, WA **October 2019 to Oct 2020**
Role: Data Engineer

Responsibilities:
- Enhanced building a centralized **Data Lake** on **AWS** Cloud utilizing primary services like **S3, EMR, Redshift, Lambda, and Glue**.
- Developed the scripts to automate the ingestion process using **Pyspark** as needed through various sources such as **API, AWS S3**, **Teradata,** and **Redshift.**
- Implemented **AWS Redshift** by Extracting, transforming, and loading data from various heterogeneous data sources and destinations.
- Deployed **Workload Management (WML)** in **Redshift** to prioritize basic dashboard queries over more complex longer running **Adhoc** queries, this allowed for a more reliable and faster reporting interface, giving sub-second query responses for basic queries.
- Build a real-time streaming pipeline utilizing **Kafka**, **Spark Streaming**, and **Redshift**.
- Worked on migrating datasets and **ETL workloads** from On-prem to **AWS Cloud**.
- Built a series of **Spark** applications and **Hive scripts** to produce various analytical datasets needed for digital marketing teams.
- Worked extensively on fine-tuning **spark** applications and providing production support to various pipelines running in **production**.
- Leverage graphing capabilities for analysing healthcare data.
- Worked on automating the infrastructure setup, launching, and termination **of EMR clusters**.
- Worked on creating **Kafka** producers using **Kafka** Producer **API** for connecting to external Rest live stream applications and producing messages to **Kafka** topic.
- Implemented ETL Migration services by developing the **AWS Lambda** functions to generate a serverless data pipeline that can be written **to AWS Glue Catalog** and queried from **AWS Athena**.
- Developed **Python script** using **Boto3** library to download files from **AWS S3** bucket and utilized Python script in **SSIS** package for **ETL** processing through SQL stored procedures.
- Responsible for creating on-demand tables on **S3** files using **Lambda** Functions and **AWS Glue** using **Python** and **Pyspark**.
- Developed robust and scalable data integration pipelines to transfer data from the s3 bucket to the Redshift database using **Python** and **AWS Glue**

- Managed metadata to ensure the quick and efficient finding of data for customer projects in **AWS Data Lake** and its complex functions like **AWS Lambda**, **and AWS Glue**.
- Created monitors, alarms, notifications, and logs for **Lambda** functions, **Glue** Jobs, and **EC2** hosts using **CloudWatch.**
- Created **Hive** external tables on top of datasets loaded in **S3** buckets and created various hive scripts to produce a series of aggregated datasets for downstream analysis.
- Developed the AWS **glue ETL** data transformation, validation, and data cleansing and catalog with crawler to get the data from **S3** and perform SQL query operations.
- Designed in creating **S3** buckets in custom policies for access management for the clients using **AWS IAM** (Identity Access Management).
- Extensive Experience in working with Cloudera (CDH4 & 5), Hortonworks Hadoop distros, and **AWS** Amazon **EMR**, to fully leverage and implement new Hadoop features.
- Created **Airflow DAGs** for Batch Processing to orchestrate **Python** data pipelines for CSV file preparation pre-ingestion, using conf to parameterize for a multitude of input files.
- Orchestrated the **Apache Airflow** to author workflows as **directed acyclic graphs (DAGs**), to visualize batch and real-time data pipelines running in production, monitor progress, and troubleshoot issues when needed.
- Enhanced the design and implementation of fully automated Continuous Integration, Continuous Delivery, Continuous Deployment pipelines, and DevOps processes for **Agile** projects **(CI/CD)**.
- Managed workload and utilization of the team. Coordinated resources and processes to achieve **Tableau** implementation plans.
- Responsible for deploying final workflows to production and maintaining and supporting production pipelines.

Environment: AWS S3, Apache Airflow, EMR (Elastic Map Reduce), Redshift, Athena, Glue, Spark, Scala, Python, Hive, Kafka, Pig, HDFS, git, DynamoDB, Lambda, Step Functions, Parquet, Tableau.

Client: Grape soft solution-Bangalore, India **November 2016 to Aug 2019**
Role: Data Engineer
Responsibilities:

- Enhanced end-to-end development of **Data Warehouses/Data Marts/Data Lakes** with **ETL** tools like **Informatic power center,** and **Big Data platform (PySpark, Hive, Hadoop Ecosystem)** environments.
- Created **Databricks notebooks** using **Spark SQL, Scala, Python**, and **Automated notebooks** using jobs.
- Well-versed in migrating data from an on-premises environment to **Azure Environment** by using **Azure Data Factory**.
- Analyzed and transformed the on-premises **SQL scripts** and designed the solution to implement using **PySpark** in **Databricks.**
- Developed data ingestion jobs using different steaming services like **Apache Kafka** into different data storage services in **Azure** and other enterprise data stores.
- Developed the scalable data pipelines **in Azure Databricks** and ingested the enrichment on the **gold layer of the data lake**.
- Built customer operator tasks in **Azure Integration Services** using Python-based data pipeline use cases.
- Experienced **ETL** resource in automating the data processing pipelines using **Azure Integration Services** and monitoring the workflows end to end.
- Extensively worked towards performance tuning/optimization of queries, contributing to a 15% - 30% improvement in the deployed code using partitioning, and clustering techniques.
- Developed templates to trigger the Azure jobs from requests and integrated the **Dataflow** with Azure for storage.
- Build the infrastructure required for **optimal extraction**, **transformation,** and **loading** of data from a wide variety of data sources using **SQL** and **Azure** 'big data' technologies.

- Built and automated data engineering **ETL pipeline** over **Snowflake DB** using **Apache Spark** and integrated data from disparate sources with Python APIs like **PySpark** and consolidated them in a data mart (Star schema).
- Experienced in managing the Terabytes of historical data stored in cloud storage like **Azure cloud Storage.**
- Worked in implementing, Building, and Deployment of **CI/CD pipelines**, managing projects often includes tracking multiple deployments across multiple pipeline stages (Dev, Test/QA staging, and production).
- Responsible for end-to-end deployment of the project that involved **Data Analysis, Data Pipelining, Data Modelling, Data Reporting, and Data documentation** as per the business needs.
- Experienced in leveraging data **visualization** tools like **Power BI** for the end customer.
- Performance analysis and fixing issues for Spark Jobs to optimize the execution time to reduce the cost of execution resources.

Environment: Apache Spark, Spark SQL, Azure Data Lake, Azure Data Bricks, Azure Blob storage, Azure Data Factory, Azure Synapse Analytics, Azure SQL Data warehouse, Azure HD Insights, Azure Functions, PySpark, SQL, Snowflake, Apache Kafka, Azure Integration Service.

Client: Broadridge Financial Solutions, Hyderabad, India October 2014 to Oct 2016
Role: Big Data Engineer

Responsibilities:
- Performance analysis and fixing issues for Spark Jobs to optimize the execution time to reduce the cost of execution resources.
- Ingested hundreds of millions of records daily from diverse data sources using **Cloudera Hadoop** cluster in the **big data Hadoop ecosystem.**
- Utilized Zookeeper for coordination and management of distributed systems.
- Managed large datasets in a columnar format within a **NoSQL** data store.
- Utilized **Kafka, Elasticsearch, and complex regex** for data ingestion, cleaning, and transformation for machine learning projects.
- Utilized AWS services such as **S3, Glue, and Athena** to build data catalogs and enable efficient data search.
- Set up **Google Cloud Platform (GCP) Dataproc clusters** for data transformation and analytics in Google Cloud Storage.
- Configured and managed **NiFi data** flow clusters for efficient data movement.
- Used **Pentaho** for building **ETL pipelines**, integrating with **Kafka, Elasticsearch, AWS S3.**
- Extracted and mapped data from various formats, including **XML, JSON, binary, and base64** encoding/decoding.
- Demonstrated expertise in the full software development cycle.
- Applied object-oriented principles to design and implement scalable and maintainable solutions.
- Worked with distributed systems, multiple-tier **Service-Oriented Architecture (SOA)**, **Java development, Scala, and Shell Scripting.**

Environment: Cloudera Distribution of Hadoop (CDH), HDFS, Hive, NoSQL, Apache Cassandra, MS SQL Server 2014, Java, Scala, Apache airflow.