Name: Srikanth Bhompally
Data Engineer
Phone: 5133027340
Email: bhompallys@gmail.com
LinkedIn ID:

Summary:

- Data Engineer with over 9 years of experience in Data Engineering, Data Warehousing, Feature Engineering, Big Data, ETL/ELT, and Business Intelligence.
- Good experience in all phases of Software Development with Agile methodology which includes User Interaction, Design & Development, Business Analysis/Modelling, Integration, Planning and testing, migration and documentation in applications using ETL pipelines and distributed applications.
- Experience in the end-end process from requirements gathering to implementation using software development methodologies such as Agile Software Development, Scrum, Test Driven Development (TDD), Data Pipeline Design, development, and implementation using Continuous Integration and Continuous Deployment (CI/CD).
- Strong experience in Amazon AWS, Google Cloud Platform and Microsoft Azure cloud services.
- Experience with Data Streaming from various sources cloud (AWS, Azure) on - premises by using the tools Spark Flume.
- Strong experience in AWS and Azure frameworks, Cloudera, Hadoop Ecosystem, Spark/PySpark/Scala, Data Bricks, Hive, Redshift, Snowflake, Relational Databases, tools like Tableau, Airflow, Presto/Athena, and Data DevOps Frameworks/Pipelines with strong Programming/Scripting skills in Python.
- Experience with Azure transformation projects and Azure architecture decision making to implement ETL and data movement solutions using Azure Data Factory (ADF) and SSIS.
- Experience in building the Orchestration on Azure Data Factory for scheduling purposes.
- Strong experience on data warehousing concepts and data modelling techniques such as dimensional data modelling, star schema modelling and snowflake modelling.
- Experience working with file structures such as CSV, JSON, parquet, and Avro file formats.
- Experience working with Azure data services, such as Azure SQL, data flow, storage and BI solution.
- Expertise on designing and developing the Big Data Analytics platforms using Big Data, Spark, Real-time streaming, Kafka, Machine Learning, and Cloud.
- Experience in using Spark Data Frames, Spark-SQL and RDD API of Spark for performing various data transformations and dataset building. Developed RESTful web Services to retrieve, transform and aggregate the data from different end points to Hadoop (HBase).
- Strong experience in data processing and transformations and actions in Spark by using Python (PySpark) language.
- Good experience on architecture and components of Spark, and efficient in working with Spark Core, Spark SQL, Spark streaming and expertise in building PySpark and Spark-Scala applications for interactive analysis, batch processing and stream processing.
- Experience in batch processing and writing programs using Apache, Spark for handling real-time analytics and real Streaming of data.
- Good knowledge in GCP cloud services like Cloud Dataflow, DataProc, CloudComposer and BigQuery.
- Good experience on Zookeeper and Kafka for monitoring and managing Hadoop jobs and used Cloudera CDH4, CDH5 for monitoring and managing Hadoop Cluster.
- Experience in Analysis, Design, Development and Big Data in Scala, Spark, Hadoop, Pig and HDFS environment.
- Good experience in AWS ECO system like amazon s3, RDS, EC2 instance and Auto Scaling, EMR, Cloud Formation template, Amazon Machine Image, S3 Glacier, Lambda, SNS, Cloud Watch, etc.
- Experience in building Machine Learning solutions using PySpark for large sets of data on Hadoop ecosystem.
- Good knowledge of tools like Snowflake, SSIS, SSAS, SSRS to design warehousing applications.
- Experience in data mapping, exploratory data analysis and logical data modelling between source and destination systems.
- Experienced on different Relational Database Management Systems like PostgreSQL, MySQL, Oracle and SQL Server.
- Experienced in developing Power BI Reports and Dashboards from multiple data sources using Data Blending.
- Hands on experience in data mining, including predictive behavior analysis, Optimization and Customer Segmentation analysis using SAS and SQL.
- Good understanding and knowledge of NoSQL databases like MongoDB, HBase and Cassandra.
- Extensive experience in development of CI/CD pipelines using azure DevOps service.
- Hands-on experience in working on high availability methodologies for Azure Cloud and Azure SQL.
- Experience in Applied Statistics, Exploratory Data Analysis and Visualization using Matplotlib, Tableau and Power BI.

Technical Skills:

Big Data Technologies	**HDFS, MapReduce, Hive, YARN, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Flume, Impala**
Languages	**Python, Scala, SQL, Unix Shell Scripting, Java, R, C/C++**
Databases	**Oracle, MySQL, SQL, Cassandra, MongoDB, PostgreSQL, HBase, Snowflake**
Operating Systems	**Windows, LINUX, UNIX, Ubuntu, Mac OS**
IDEs / Reporting Tools/Design	**MS Visual Studio, Eclipse, PyCharm, Tableau, Power BI**
Version Control	**Git, Bitbucket, GitHub**
CI/CD Tools	**Kubernetes, Docker, Terraform, Jenkins.**
Cloud Platforms	**Amazon Web Services AWS, Microsoft Azure, Google Cloud Platform GCP**
Monitoring Tools	**ANT, Maven**

Work Experience:

Client: Amazon **Apr 2022- Present**
AWS Data Engineer
Responsibilities:

- Utilized AWS Glue to design and implement scalable and efficient ETL (Extract, Transform, Load) processes for ensuring smooth data flow and integration across various systems.
- Leveraged Hadoop ecosystem including HDFS and MapReduce to store and process structured and unstructured data efficiently, enabling comprehensive analysis and insights generation.
- Implemented and optimized data workflows on Databricks platform, leveraging Spark for high-speed data processing and analytics, ensuring timely delivery of actionable insights to stakeholders.
- Developed data integration solutions using Talend, enabling seamless connectivity between disparate data sources and destinations while ensuring data quality and integrity.
- Engineered data pipelines on Impala and Hive for fast SQL queries over large datasets, facilitating ad-hoc analysis and reporting requirements.
- Implemented data ingestion and processing workflows using Apache Airflow and Hadoop on AWS, managing and optimizing data pipelines for scalable and efficient big data processing.
- Automated deployment processes using Jenkins for continuous integration and continuous deployment (CI/CD), enhancing development efficiency and reducing time to market for data solutions.
- Orchestrated data flows and event processing using Nifi, ensuring reliable data ingestion and transformation from various sources into the data lake environment.
- Developed data processing applications in Scala to leverage the functional programming paradigm and improve code maintainability and scalability.
- Implemented NoSQL databases such as MongoDB and Cassandra for storing and querying semi-structured and unstructured data, supporting flexible data modeling and analysis.
- Utilized Python for scripting and automation tasks, enabling efficient data manipulation, cleansing, and transformation processes.
- Designed and optimized data workflows using Pig and Sqoop for efficient data transfer between Hadoop and relational databases, ensuring data consistency and integrity.
- Integrated Hibernate and Spring frameworks into data processing applications, facilitating seamless interaction with relational databases and improving code modularity and reusability.
- Scheduled and coordinated workflow executions using Oozie, ensuring timely and reliable execution of data processing tasks and job dependencies.
- Leveraged various AWS services including EC2, S3, and Autoscaling for provisioning and scaling compute and storage resources based on workload demands, optimizing cost and performance.
- Developed data processing applications using Scala on AWS infrastructure, leveraging cloud-native services for efficient and scalable data processing.

- Implemented full-text search capabilities using Elasticsearch, enabling fast and accurate retrieval of data for analysis and reporting.
- Designed and optimized data storage and retrieval processes using DynamoDB, ensuring low-latency access to critical data for real-time applications.
- Developed and maintained UNIX shell scripts for automating system administration tasks and managing data processing workflows in a Unix environment.
- Optimized query performance and resource utilization using TEZs for efficient execution of Hive and Pig jobs on Hadoop clusters.
- Designed and implemented data warehousing solutions using AWS Redshift, enabling high-performance analytics and reporting on data at scale.
- Orchestrated data movement and transformation using AWS Data Pipelines, ensuring reliable and efficient data processing workflows across various AWS services.
- Successfully led migrated data processes to AWS Redshift, reducing overall infrastructure costs by 15%.

Environment: AWS Redshift, ETL, MapReduce, DynamoDB, Hive, PIG, Hadoop, UNIX, AWS EC2, AWS S3, Hibernate, Oozie, PIG, Sqoop, MongoDB, Cassandra, Impala, SQL, AWS Glue, CI/CD, Jenkins, Apache Airflow, AWS Autoscaling, Scala.

Client: JPMG **Jan 2021 – March 2022**
Azure Data Engineer
Responsibilities:

- Made logical data models and physical data models that capture current state/future state data elements and data flows.
- Participated with the project team and business analysts to define project requirements in JRD meetings.
- Utilized Python for recurring reports automation and visualized them on BI platform.
- Utilized abilities in maintaining data integrity by developing constraints and DML triggers.
- Maintained a document of all the sources utilized, new tables/views created, and data gaps in the existing data system.
- Created the data pipeline from source to database like Snowflake/PostgreSQL.
- Worked on the automation, innovation to reduce the manual efforts.
- Worked on the SQL operations using data frame in the spark environment.
- Worked on the tokenization for the file level, field level using PySpark.
- Developed and implemented data cleansing, data security, data profiling and data monitoring processes.
- Analyzed the existing data model for optimizations and data for inconsistencies.
- Mapped data sources and targets for data movement and analyzed it to ensure appropriate integrity and quality.
- Generated periodic reports based on the statistical analysis of the data from various time frame and division.
- Developed data mapping, data governance, and transformation and cleansing rules involving OLTP and OLAP.
- Involved heavily on data profiling and data cleansing utilizing lookup, fuzzy lookup, and fuzzy grouping per company data standards,
- Utilized Azure Data Factory to perform Azure-SSIS integration and deployed SSIS packages to cloud.
- Handled structured and unstructured datasets.
- Built high quality, reliable, and consistent sound systems that are aligned and scale with our data business needs.
- Recreated existing application logic and functionality in the Azure Data Factory, Azure SQL Database and Azure SQL Data Warehouse environment.
- Performed evaluation of on-premises SQL database to Azure SQL.
- Implemented Azure Data Factory operations and deployed into Azure for moving data from on-premise into cloud.
- Designed and implemented end-to-end data solutions (storage, integration, processing, and visualization) in Azure.
- Architected & implemented medium to large scale BI solutions on Azure DevOps using Azure Data Factory.
- Worked closely with the Enterprise Data Warehouse team and Business Intelligence Architecture team to understand repository objects that support the business requirement and process.
- Analyzed the source data and worked with business users and developers to develop the data model.
- Outlined security policies for Power BI reports and linked folders and reports with data sources.
- Transformed data in Azure Data Factory with the ADF Transformations.
- Conceptualized the ETL process to pull and populate data from OLTP sources to the data mart.
- Administered Power BI gateways by connecting datasets and updated the gateway with missing data sources.
- Worked on Azure Data Lake storage and loading data into Azure SQL Data Warehouse.
- Architected & implemented medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Factory, Azure SQL Server & Azure SQL DW).

- Implemented Azure Data Factory operations and deployed into Azure for moving data from on-premise into cloud.

Environment: Power BI, Snowflake, PostgreSQL, Python, Spark, SQL, PySpark, SSIS, Azure SQL, Azure Data Factory, Azure Data Lake, Azure SQL DW.

Client: Hanover Insurance, MA **Jan 2019 – Dec 2020**
GCP Data Engineer
Responsibilities:

- Worked on cloud environment GCP services. Designed and developed batch and streaming pipelines using GCP services for different clients.
- Involved in the design and building of project right from the scratch.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDD, Scala, and Python. Worked in Agile development environment in sprint cycles of two weeks by dividing and organizing tasks. Participated in daily scrum and other design related meetings.
- Collaborated with product teams, data analysts and data scientists to design and built data-forward solutions.
- Key contributor in building the complete workflows, triggers, PySpark jobs, crawlers, Lambda functions.
- ETL pipelines in and out of data warehouse using combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.
- Wrote various data normalization jobs for new data ingested into Redshift.
- Designed and developed batch and streaming pipelines using GCP services like Cloud Storage, BigQuery, Compute Engine, Cloud Composer, Cloud Monitoring and cloud Function.
- Migrated the on-perm solutions to GCP cloud, created Data Lake in cloud storage and data warehouse in Big Query.
- Automated and Orchestrated pipelines using Google composer, Created Airflow DAGS.
- Process involved Design, Development, Build, Testing, Implementation, and support till there were no further issues.

Environment: ETL, Snowflake, Python, Apache Airflow, Hive, Scala, SQL, Spark, PySpark, Cloud Storage, GCP BigQuery, Compute Engine, Cloud Composer, Cloud Function, Google composer.

Client: Frontier Communications Corporation, TX **June 2017 – Dec 2018**
Hadoop Developer
Responsibilities:

- Data Ingestion implemented using SQOOP, SPARK, and loading data from various RDBMS, CSV, and XML files.
- Data cleansing and transformation tasks are handled using SPARK using SCALA, and HIVE.
- Data Consolidation was implemented using SPARK, and HIVE to generate data in the required formats by applying various ETL tasks for data repair, massaging data to identify sources for audit purposes, data filtering, and storing back to HDFS.
- Responsible for design development of Spark SQL Scripts based on Functional Specifications.
- Exploring with the Spark improving the Performance and Optimization of the existing algorithms in Hadoop.
- ETL development to normalize this data and publish it in IMPALA.
- Involved in converting Hive/SQL queries into Spark RDD using Scala.
- Responsible for Job management using Fair scheduler and Developed Job Processing scripts using Oozie Workflow.
- Responsible for Performance Tuning of Spark Applications for setting the right Batch Interval time, correct level of Parallelism, and Memory tuning.
- Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, and Pair RDD's.
- Importing and exporting data into HDFS and HIVE, PIG using Sqoop.
- Involved in creating Hive Tables, loading with data, and writing Hive queries, which will invoke and run MapReduce jobs in the backend.
- Implemented the workflows using the Apache Oozie framework to automate tasks.
- Worked with No SQL databases like HBase. Creating HBase tables to load large sets of semi-structured data coming from various sources.
- Worked with different file formats such as Text, Sequence files, Avro, ORC, and Parquet.
- Responsible to manage data coming from different sources.
- Responsible for loading and transforming large sets of structured, semi-structured, and unstructured data.
- Analyzed large amounts of data sets to determine the optimal way to aggregate and report on them.

Environment: Sqoop, Spark, Apache Oozie, Scala, IMPALA, HBase, HDFS, Hive, PIG, MapReduce, SQL, Spark-SQL.

Client: Rivian Automotive, CA **Feb 2014 – July 2016**
Data Consultant
Responsibilities:

- Involved in database design, data modeling, T-SQL development; developed and implemented ETL packages using SSIS and complex reports.
- Participated in the creation of data standards, policies and procedure to ensure future data quality.
- Analyzed multiple user requirements and development specifications for OLAP applications in JAD meetings.
- Analyzed user requirements, understood business rules in order to build enterprise-wide data quality environment.
- Maintained data integrity with the creation of constraints and DML triggers.
- Wrote Python script for manipulating and looping through different user defined objects.
- Implemented Agile for frequent changes to client requirements and following parallel development and testing.
- Participated in analyzing source databases and creating data mapping documents for the data migration process.
- Designed and developed complex ETL mappings to migrate data from the source and load it into the staging and Warehouse Table.
- Created SQL scripts to fetch reports from table.
- Developed a parent-child architecture on database objects including SPs, UDFs, and tables.
- Created and maintained SSIS package (ETL) to extract and transform data, and to load into staging database.
- Utilized tasks such as script task, FTP task and Execute SQL task in SSIS.
- Enabled package executions from the latest point of failure with checkpoints in SSIS.
- Implemented error handling utilizing try-catch block and error functions in batches.
- Worked on all activities related to the development, implementation, administration and support of ETL processes for large-scale data warehouses using SSIS.
- Implemented dynamic data masking in Azure SQL Database and Azure Synapse Analytics with different masking functions and data types using the Azure portal and T-SQL commands.
- Designed periodic reports based on the statistical analysis of the data from different timeframe in SSRS.
- Developed, designed and deployed reports through SSRS and Excel, per business requirements.

Environment: Azure SQL Database, Azure Synapse Analytics, T-SQL, SSRS, MS Excel, SSIS, ETL, Python.