**Varun**

**Mail:** [**bvkumar2899@gmail.com**](mailto:bvkumar2899@gmail.com)

**Phone: (704) 777-5958**

**PROFESSIONAL SUMMARY:**

* Over 10+ years of IT experience in a variety of industries, which includes hands on experience in **Big Data, Hadoop, and Java development.**
* Expertise with the tools in Hadoop Ecosystem including **Spark, Hive, HDFS, MapReduce, Sqoop, Kafka, Yarn, Oozie,** and **HBase.**
* Excellent knowledge on Distributed components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming paradigm.**
* Experience automating data engineering pipelines utilizing proper standards and best practices (right partitioning, right file formats, incremental loads by maintaining previous state etc.,)
* Experience in designing and developing production ready data processing applications in **Spark** using **Scala/Python.**
* Experience working with **Snowflake** Multi cluster and virtual warehouses in Snowflake.
* Strong experience creating the most efficient Spark applications for performing various kinds of data transformations like data cleansing, de-normalization, various kinds of joins, data aggregation.
* Experience fine-tuning **Spark** applications utilizing various concepts like Broadcasting, increasing shuffle parallelism, caching/persisting Data Frames, sizing executors appropriately to utilize the available resources in the cluster effectively etc.,
* Strong experience in writing applications using python using different libraries like **Pandas, NumPy, SciPy, Matplotlib etc.**
* Good Knowledge in productionizing Machine Learning pipelines (Featurization, Learning, Scoring, Evaluation) primarily using **Spark ML libraries.**
* Good exposure with **Agile software** development process.
* Experience in manipulating/analyzing large datasets and finding patterns and insights within structured and unstructured data.
* Strong experience on Hadoop distributions like **Cloudera, Hortonworks, AWS, and Azure Databricks**.
* Good understanding of **NoSQL** **databases** and hands-on work experience in writing applications on NoSQL databases like **HBase, Cassandra and MongoDB.**
* Experienced in writing complex **MapReduce** programs that work with different file formats like **Text, Sequence, Xml, parquet, and Avro.**
* Experience in Oozie and workflow scheduler to manage Hadoop jobs by Direct Acyclic Graph (DAG) of actions with control flows.
* Experience in migrating the data using Sqoop from HDFS to Relational Database System and vice-versa.
* Extensive Experience on importing and exporting data using stream processing platforms like Flume and Kafka.
* Excellent Java development skills using **J2EE, J2SE, Servlets, JSP, EJB, JDBC, SOAP and RESTful web services.**
* Experience in database design using **PL/SQL** to write Stored Procedures, Functions, Triggers, and strong experience in writing complex queries for Oracle.
* Software development involving cloud computing platforms like Amazon Web Services (AWS), and Google Cloud (GCP).
* Experienced in working with Amazon Web Services (AWS) using S3, EMR, Redshift, Athena, Glue Meta store etc.,
* Experience in **GCP, Big Query, GCS bucket, G - cloud** function, cloud dataflow, Data Proc, Storage, Composer and Stack driver.
* Strong experience in Object-Oriented Design, Analysis, Development, Testing, and Maintenance.
* Experienced in using agile approaches, including Extreme Programming, Test-Driven Development and Agile Scrum.
* Worked in large and small teams for systems requirement, design & development.
* Key participant in all phases of software development life cycle with **Analysis, Design, Development**, **Integration**, **Implementation**, **Debugging**, and Testing of Software Applications in client server environment, Object Oriented Experience in using various IDEs Eclipse, IntelliJ, and repositories SVN and Git.
* Experience in complete project life cycle (design, development, testing and implementation) of Client Server and Web applications.
* Experience of using build tools **SBT, Maven.**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Operating Systems** | Linux, Windows XP/7/8/10, Mac. |
| **Programming languages** | Core Java, Spring Boot, R, Scala, Terraform. |
| Big Data Technologies | HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, and Hbase, Spark, Spark Streaming, Yarn, Zookeeper, Kafka, Attunity, ETL. (Nifi, Talend etc.) |
| **Scripting/Web Languages** | Ab-Initio, HTML5, CSS3, XML, SQL, Shell/Unix, Perl, Python. |
| **Databases** | Cassandra, HBASE, MongoDB, Oracle, MS SQL, MY SQL, Teradata. |
| **Utilities/Tools** | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven, Alteryx, Visio, Jenkins, Jira, Intellij. |
| **Cloud Services** | AWS (EC2, S3, EMR, RDS, Lambda, CloudWatch, FSx, Auto scaling, Redshift, Cloud Formation, Glue etc.) Azure (Databricks, Data Lake, Data Factory, Data Storage, HDInsight, SQL etc.) |
| **Data Visualization Tolls** | Tableau, SSRS, Cloud Health. |
| **Software Life Cycle** | SDLC, Waterfall and Agile models. |

**WORK EXPERIENCE:**

**Client: CVS Health - Woonsocket, RI Jan 2022 – Till date**

**Role: Senior Big Data Engineer**

**Responsibilities:**

* Create **Spark** Vectorized panda user defined functions for data manipulation and wrangling.
* Transfer data in logical stages from System of records to raw zone, refined zone and produce zone for easy translation and de normalization.
* Setting up **Azure infrastructure** like storage accounts, integration runtime, service principal id, app registrations to enable scalable and optimized utilization of business user analytical requirements in **Azure**.
* Configured **Spark streaming** to get ongoing information from the **Kafka** and store the stream information to **HDFS**.
* Deployed the Big Data Hadoop application using Talend on cloud AWS (Amazon Web Services) and also on Microsoft Azure.
* Extracted and updated the data into **HDFS** using **Sqoop** import and export.
* Developed **HIVE UDFs** to incorporate external business logic into **Hive** script and developed join data set scripts using **HIVE** join operations.
* Involved in creating **HDInsight** cluster in **Microsoft Azure Portal** also created **Eventshub** and **Azure SQL Databases.**
* Designed and implemented **HL7 interfaces** between various healthcare information systems, including **EHR, LIS & RIS.**
* Worked on a clustered **Hadoop** for **Windows Azure** using **HDInsight** and **Hortonworks** Data Platform for Windows.
* Experience in Hadoop Big Data Integration with Talend ETL on performing data extract, loading and transformation process.
* Assessed the **Netezza** **environment** for implementation of the **ETL** solutions.
* Created **Netezza** SQL scripts to test the table loaded correctly.
* Implemented and managed for Devops infrastructure architecture, Terraform, Jenkins, Puppet and Ansible implementation, Responsible for CI infrastructure and CD infrastructure and process and deployment strategy.
* Writing **PySpark** and spark **Sql transformation** in **Azure Databricks** to perform complex transformations for business rule implementation.
* Implemented **Kafka** producers create custom partitions, configured brokers and implemented High level consumers to implement data platform.
* Used Bash and Python, including Boto3 to supplement automation provided by Ansible and Terraform for tasks such as encrypting EBS volumes backing AMIs and scheduling Lambda functions for routine AWS tasks
* Conducted HL7 testing and troubleshooting to ensure accurate and timely data exchange.
* Create and maintain optimal data pipeline architecture in cloud **Microsoft Azure** using **Data Factory** and **Azure Databricks**
* Create Self Service reporting in **Azure Data Lake Store Gen2** using an **ELT approach.**
* Built real time pipeline for streaming data using **Eventshub/Microsoft Azure Queue** and **Spark** **streaming**.
* Responsible to manage data coming from different sources through **Kafka**.
* Created Talend Spark jobs, which collects data from regular relational database, and load the data in to HBase
* Built the **logical and physical data model for snowflake** as per the changes required.
* Installed **Kafka** Producer on different severs and Scheduled to produce data for every 10 seconds Implemented Data Quality in **ETL Tool Talend** and having good knowledge in Data Warehousing Developed **Apache Spark** applications by using spark for data processing from various streaming sources.
* Developed **Abinitio** code for to process **XML** files as well as database retrievals in an agile environment.
* Designed, developed, and delivered human resources data integration to multiple external systems using **Ab-Initio ETL.**
* Worked with the **Spark** for improving performance and optimization of the existing algorithms in **Hadoop** using **Spark Context, Spark-SQL, PySpark, Impala,** **Tealeaf, Pair RDD's, DevOps, Spark YARN**
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value using **Azure Data Factory**
* Implement IOT streaming with Databricks Delta tables and Delta lake to enable ACID transaction logging
* Exposed transformed data in **Azure Spark Databricks** platform to parquet formats for efficient data storage
* Developed custom **HL7 interfaces** for client using interface engines such **as Cloverleaf and Rhapsody**.
* Created sample flows in Talend, Stream sets with custom coded jars and analyzed the performance of Stream sets and Kafka steams.
* Creating Data factory pipelines that can bulk copy multiple tables at once from relational database to **Azure data lake gen2**
* Create custom logging framework for **ELT** pipeline logging using Append variables in Data factory
* Developed various **Oracle SQL scripts, PL/SQL packages**, procedures, functions, and java code for data
* Enabling monitoring and azure log analytics to alert support team on usage and stats of the daily runs
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value using **Azure Data Factory**
* Consulting on **Snowflake data platform solution architecture**, design, development, and deployment focused to bring the data driven culture across the enterprises.
* Developed and designed **data integration** and **migration solutions** in **Azure**.
* Kept our data separated and secure across national boundaries through multiple data centers and regions.
* Implement Continuous integration/continuous development best practice using **Azure DevOps**, ensuring code versioning.
* Collaborated with **cross-functional teams** to ensure successful implementation of healthcare information systems.
* Migrate Oracle Datawarehouse to Big Data Platform using **Attunity.**
* Strong Knowledge on architecture and components of Tealeaf, and efficient in working with **Spark Core,** **SparkSQL.** Designed and developed RDD Seeds using Scala and Cascading. Streaming data to **Spark** streaming using **Kafka**
* Worked with **PowerShell** and **UNIX** **scripts** for file transfer, emailing and other file related tasks.
* Good Exposure on **Map Reduce** programming using Java, PIG Latin Scripting and Distributed Application and **HDFS**.
* Used **Airflow** to monitor and schedule the work
* Utilized **Ansible playbook** for code pipeline deployment
* Worked with **NoSQL databases** like **HBase** in creating **HBase** tables to load large sets of semi-structured data coming from various sources
* Conducted **ETL** **development** in the **Netezza** environment using standard design methodologies.
* Delivered de normalized data for **Power BI** consumers for modeling and visualization from the produced layer in **Data lake**
* Worked in a SAFE (Scaled Agile Framework) team with daily standups, sprint planning, quarterly planning.

**Environment**:  Hadoop, Spark, MapReduce, Kafka, Scala, JAVA, Azure Data Lake Gen2, Azure Data Factory, PySpark, Databricks, Netezza, Azure DevOps, Agile, Power BI, Python, R, PL/SQL, Oracle 12c, SQL, No SQL, HBase, Scaled Agile team environment

**Client: NewYork State - Albany, NY Oct 2020 – Nov 2021**

**Role: Senior Data Engineer / Big Data Developer.**

**Responsibilities:**

* Involved in the development of use case documentation, requirement analysis, and project documentation.
* Developed and maintained Web applications as defined by the Project Lead.
* Developed and designed an **API (RESTful Web Service)** for the company’s website.
* Developed and designed e-mail marketing campaigns using **HTML** and **CSS.**
* Excellent Data analysis skills
* Extensive experience in Amazon Web Services (AWS) Cloud services such as **EC2, VPC, S3, IAM, EBS, RDS, ELB, VPC, Route53, Ops Works, DynamoDB, Autoscaling, CloudFront, CloudTrail, CloudWatch, CloudFormation, Elastic Beanstalk, AWS SNS, AWS SQS, AWS SES, AWS SWF & AWS Direct Connect.**
* Deployed microservices into **AWS - EC2.**
* You may develop and maintain **ETL** (Extract, Transform, Load) pipelines to extract metadata from various sources and load it into **EDC**.
* Designing, developing, deploying, and supporting computing applications in a large-scale, complex environment.
* Consumed **XML** messages using **Kafka** and processed the xml file using Spark Streaming to capture UI updates.
* Comfortable in configuring and using multiple operating systems **(Mac/Windows/\*nix)**
* Maintained customers relationship management databases **(MySQL / PostgreSQL)**
* Developed server-based web traffic statistical analysis tool using **Flask, Pandas.**
* Used **MS Visio** for creating business process diagrams.
* Ability to communicate at various levels within large organizations
* Migrated an existing on premises application to **AWS**.
* AWS Athena to query the data and do the data validation between **GCP** and **AWS**.
* Designed various tables required for the project in Oracle database and used Stored Procedures in the application.
* Used **SQL** to create, update and manipulate tables.
* Used **AWS** **Cloud** with Infrastructure Provisioning / Configuration.
* Involved in impact analysis of Change requests and Bug fixes.

**Environment:** Python, Matplotlib, Pandas, OpenCV, NumPy, Seaborn, TensorFlow, Keras, Sci-kit-learn, NLTK, AWS S3, RDS, PySpark, Tableau, QlikView, Hive, MYSQL, informatica EDC, Sqoop, HDFS, MongoDB, Bootstrap, CSS, JavaScript

**Client: DTCC - Coppell, TX March 2019 – Sept 2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Developed efficient MapReduce programs for filtering out the unstructured data and developed multiple **MapReduce** jobs to perform data cleaning and pre-processing on **Hortonworks**.
* Used **Spark Streaming** APIs to perform transformations and actions on the fly for building common learner data model which gets the data from **Kafka** in near real time and persist it to **Cassandra**.
* 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**
* 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.**
* Loaded data from different source (database & files) into Hive using Talend tool.
* Responsible in development of **Spark Cassandra** connector to load data from flat file to Cassandra for analysis.
* Develop stored procedures in snowflake and use in Talend for loading dimensions and facts.
* Used **Hortonworks** Apache Falcon for data management and pipeline process in the Hadoop cluster.
* Used **impala** to query the data into the publish layers where all the other teams or business users can access for faster processing.
* Proficient with container systems like Docker and container orchestration like EC2 Container Service, Kubernetes, worked with Terraform.
* Automated resulting scripts and workflow using **Apache Airflow** and **shell scripting** to ensure daily execution in production.
* Building Real-time Data Pipelines with **Kafka** Connect and **Spark Streaming**.
* Developed **Scala** scripts using both **Data frames/SQL/Data** sets and RDD/MapReduce in **Spark** for **Data Aggregation,** queries and writing data back into **OLTP** system through **Sqoop**
* Developed **Hive queries** to pre-process the data required for running the business process
* Developed a **Spark** job in **Java** which indexes data into **ElasticSearch** from external **Hive tables** which are in **HDFS**.
* Evaluate **Snowflake design** considerations for any change in the application.
* Install and Configure **Apache Airflow** for S3 bucket and snowflake data warehouse and created **dags** to run the **Airflow**.
* Push data as delimited files into HDFS using Talend Big data studio.
* Used **Impala** and Presto for querying the datasets.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of **Spark** using **Scala**.
* Developed Automation Regressing Scripts for validation of ETL process between multiple databases like **AWS Redshift, Oracle, Mongo DB, T-SQL**, and **SQL Server** using **Python.**
* Got chance working on **Apache NiFi** like executing **Spark script**, **Sqoop** scripts through **NiFi**, worked on creating scatter and gather pattern in **NiFi**, ingesting data from Postgres to HDFS, Fetching Hive metadata and storing in HDFS, created a custom **NiFi** processor for filtering text from Flow files etc.
* Cluster coordination services through **Zookeeper**.
* Instantiated, created, and maintained CI/CD (continuous integration & deployment) pipelines and apply automation to environments and applications. Worked on various automation tools like GIT, Terraform, Ansible.
* Extract Real time feed using **Kafka** and **Spark Streaming** and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in **HDFS**.
* Experienced in writing real-time processing and core jobs using **Spark Streaming** with **Kafka** as a data pipe-line system.
* Used **Spark** and **Spark-SQL** to read the parquet data and create the tables in hive using the **Scala API**.
* Developed multiple **Kafka** Producers and Consumers as per the software requirement specifications.
* Supporting Continuous storage in AWS using **Elastic Block Storage, S3, Glacier**. Created Volumes and configured Snapshots for **EC2 instances**
* Expert with Azure Automation Powershell, Azure Resource Manager Templates, Terraform.
* Used **Kafka** and **Kafka brokers**, initiated the **spark** context and processed live streaming information with RDD and Used Kafka to load data into **HDFS** and **NoSQL** databases.
* Developed code in **Java** which creates mapping in **ElasticSearch** even before data is indexed into.
* Using **Spark-Streaming** APIs to perform transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra
* Developed end to end data processing pipelines that begin with receiving data using distributed messaging systems **Kafka** for persisting data into **Cassandra**.
* Implemented Data Interface to get information of customers using Rest API and Pre-Process data using **MapReduce** and store into HDFS (**Hortonworks**).
* Maintained **ELK** (Elastic Search, Logstash, Kibana) and Wrote Spark scripts using **Scala shell.**
* Worked in **AWS** environment for development and deployment of custom **Hadoop applications.**
* Developed shell scripts to generate the**hive** create statements from the data and load data to the table.
* Involved in writing custom **Map-Reduce** programs using java API for data processing.
* The **Hive** tables are created as per requirement were Internal or External tables defined with appropriate static, dynamic partitions and bucketing, intended for efficiency.
* Developed **Hive** queries for the analysts by loading and transforming large sets of structured, semi structured data using **hive**.

**Environment**: HDP, Hadoop, AWS, EC2, S3 Bucket, Redshift, Cassandra, Hive, HDFS, Spark, Spark-SQL, Spark-Streaming, Scala, KAFKA, Hortonworks, Map Reduce, Apache Nifi, Airflow, Impala, Zookeeper, ELK, Sqoop, Java Oracle 12c, Sql Server, T-SQL, MongoDB, Hbase, Python and Agile Methodologies.

**Client: Nielsen Corporation - Tampa, FL Jan 2018 – Feb 2019**

**Role: Data Engineer/ETL Developer**

**Responsibilities:**

* Worked as ETL Developer/Informatica developer.
* Worked as data migration from source to target and prepare the unit test case to match the result as per the business document.
* Prepare the design documents and convert the business requirements to technical documents.
* Provided extremely innovative and quick solutions for the bug fix in Informatica Mappings.
* Moving the issues to maintenance for a fix with complete code analysis and suggestions for code fix.
* Created and configured Sessions in Informatica Workflow Manager for loading data into Data Mart tables from various heterogeneous database sources like Flat Files, relational Data Bases.
* Created Mappings for Slowly Changing Dimension tables.
* Worked as SPOC for bug fixes for a few months.
* Used the version control to maintain the mapping versions.
* Co-ordinate with the **QA** **team** in system testing.
* Worked on Test and **UAT** **bug fixes.**
* Mounted Local directory file path to **Amazon S3** using S3fs fuse, to have KMS encryption enabled on the data reflecting in S3 buckets.
* Designed and implemented **ETL pipelines** on **S3** parquet files on data lake using AWS Glue.
* Migrated the data from **Amazon Redshift** data warehouse to Snowflake.
* Involved in code migration of quality monitoring tool from AWS EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
* Used **AWS** **glue** catalog with crawler to get the data from S3 and perform **SQL** **query** operations and **JSON** schema to define table and column mapping from S3 data to Redshift.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.

**Environment**: ETL pipelines, Informatica, Amazon S3, Redshift, AWS Glue, AWS Lambda, Snowflake, JSON, UAT

**System Soft Technologies - India March 2014 – June 2017**

**Role: Data Analyst/ETL Developer**

**Responsibilities:**

* Created the Design level document from the business specifications provided by the client and SMEs.
* Coordinating with the business analyst to get a clear understanding of the business and meet their expectation.
* Worked on stored procedures per the business requirements.
* Worked on **Talend** to generate the reports in form of **CSV files.**
* Created the unit test case document for every code migration to the next environment.
* Created and configured Sessions in Informatica Workflow Manager for loading data into tables from various heterogeneous database sources like Flat Files, relational Databases.
* Used the tortoise **SVN** **version** control to maintain the mapping versions.
* Co-ordinate with the **QA** and **UAT** team for any issues and provided the fixes.
* Automated strategic business planning process by deploying centralized Reporting/ Data Catalogue from source to target data.
* Optimized query performance and data quality testing to ensure action ability and reliability of data.
* Implemented data management capabilities required for the future and data quality systems and processes
* Updated/Managed data and queries using **SQL** and retrieve data related to equipment’s performance and quality reports.
* Automated migrating reports from **AWS** using **EC2** in Alteryx to Tableau sever and successfully archived report
* Generated ad-hoc **SQL** queries using joins, database connections and transformation rules to fetch data from legacyDB2 and SQL Server database systems.
* Developing visualization in **Tableau** and **Looker** for better understanding of the customer experience
* Analytical skills in order to relate specific customer asset situations to individual products
* Developed **Data Mapping, Data Governance, Transformation** and Cleansing rules for the Master Data Management Architecture involving **OLTP** and **OLAP**.
* Performed data analysis and data profiling using complex SQL on various sources systems and answered complex business questions by providing data to business users.

**Environment**: ETL, Informatica, Talend, SQL, AWS, EC2, Tableau server, OLTP, OLAP, QA

**Zeta Global - India August 2012 – Feb 2014**

**Role: Data Analyst**

**Responsibilities:**

* Document the complete process flow to describe program development, logic, testing, and implementation, application integration, coding.
* Recommended structural changes and enhancements to systems and Databases.
* Conducted Design reviews and technical reviews with other project stakeholders.
* Was a part of the complete life cycle of the project from the requirements to the production support
* Created test plan documents for all back-end database modules
* Used **MS Excel, MS Access,** and **SQL** to write and run various queries.
* Worked extensively on creating tables, views, and **SQL queries** in **MS SQL Server**.
* Worked with internal architects and assisting in the development of current and target state data architectures.
* Created Hive target tables to hold the data after all the **PIG ETL** operations using **HQL**.
* Analyzed the data by performing Hive queries and running Pig scripts to know user behaviour.
* Worked on migrating existing mainframe data and reporting feeds to Hadoop.
* Integrated Hadoop with **Tableau** and **SAS** analytics to provide end users analytical reports
* Worked with Hadoop in Stand-alone, pseudo and distributed modes.
* Coordinate with the business users in providing appropriate, effective, and efficient way to design the new reporting needs based on the user with the existing functionality.
* Remain knowledgeable in all areas of business operations to identify systems needs and requirements.

**Environment**: SQL, SQL Server, MS Office, and MS Visio