**Vinod**

**Big Data Developer**

[**Scott.d@lavsun.com**](mailto:Scott.d@lavsun.com)

**214-699-6298**

**PROFESSIONAL SUMMARY:**

* **Over 8 years** of experience in building big data applications, data pipelines, creating data lakes to manage structured and semi-structured data and workflow implementations using Big data Ecosystems like **Hadoop, Spark, Kafka,** etc.
* Expertise in developing applications using **Java** and **Python.**
* Expertise in working with Hadoop Distributions like **EMR, Hortonworks,** and **Cloudera**.
* Extensive experience in big data technologies like **HDFS, MapReduce, HBase, Spark, Yarn, Kafka, Pig, Hive, Sqoop, Storm, Oozie, Zookeeper, Impala, NiFi,** and **Flume.**
* In-depth understanding of **Hadoop Architecture** and its various components such as **Resource Manager, Node Manager, Applications Master, Name Node, Data Node concepts.**
* Worked extensively in Real-time streaming data using **Storm**, **Spark-Streaming,** and **Kafka.**
* Experienced in integrating **Kafka** with **Spark streaming** for high-speed data processing.
* Expertise in writing **DDLs** and **DMLs** scripts for analytics applications in **MySQL**, **Hive, and RDS.**
* Experienced in **Python** development for various ETLapplications as well as working with **python libraries** like **Matplotlib, Numpy, Scipy,** and **Pandas** for **data analysis.**
* Experienced in developing Hadoop applications on **SPARK** using **SCALA** as a functional and object-oriented programming.
* Expertise in working with **AWS** cloud services like **EC2, S3, Redshift, EMR, Lambda, DynamoDB, RDS, SNS, SQS, Glue, Data Pipeline, Athena** for big data development.
* Expertise in developing **Pyspark** applications using various abstractions like **RDD, Dataframes, and Datasets**.
* Expertise in working with **Hive** optimization techniques like **Partitioning, Bucketing, vectorizations** and **Map side-joins, Bucket-Map Join, skew joins,** and creating **Indexes**.
* Expertise in debugging and tuning failed and long-running **Spark applications** using various optimization techniques for executor tuning, **memory management, garbage collection, Serialization**, broadcast variable, and persisting methods assuring the optimal performance of applications.
* Expertise in **ETL** process using **python, RDS, Data Lake** to extract, transform, and load a large amount of data and implementing it on **AWS** for **data warehousing** and data migration purposes.
* Experience working with batch processing and operational data sources and migration of data from traditional databases to **Hadoop** and **NoSQL** databases.
* Experienced with different file formats like **Parquet, ORC, CSV, Text, XML, JSON,** and **Avro** files.
* Expertise in **Python scripting** and **bash scripting**.
* Excellent technical and analytical skills with a clear understanding of the design goals of ER modeling for **OLTP** and dimension modeling for **OLAP.**
* Experience in configuring the **Zookeeper** to coordinate the servers in clusters, job scheduling for Spark Jobs, and to maintain data consistency.
* Expertise in data ingestion using **Flume, Sqoop,** and **Nifi.**
* Experience in orchestrating workflows using **Oozie** and **Airflow**.
* Good Knowledge in making and keeping up profoundly versatile and **fault-tolerant** Infrastructure in **AWS** environment spanning over different availability zones.
* Passionate about gleaning insightful information from massive datasets and developing a culture of sound, data-driven decision making.
* I am a good **team player** who likes to take initiative and seek out new challenges.
* Excellent communication skills can work in a **fast-paced multitasking** environment both independently and in a collaborative team, a self-motivated enthusiastic learner.
* Involved in all the phases of Software Development Life Cycle (**Requirements Analysis, Design, Development, Testing, Deployment, and Support**) and **Agile methodologies**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Technologies** | HDFS, Apache NiFI, Map Reduce, Sqoop, Flume, Pig, Hive, HBase, Oozie, Impala, Zookeeper, Storm, Spark, and Kafka. |
| **Hadoop Distributions** | Cloudera, MapR, and EMR |
| **Languages** | Python, SQL, Scala, HiveQL, Shell Scripting, Unix, C, C++ |
| **No SQL Databases** | HBase, Cassandra, and MongoDB |
| **packages** | Pandas, Numpy, Boto3, Matplotlib |
| **AWS Services** | EC2, ELB, EMR, Redshift, RDS, IAM, S3, AWS Lambda, EBS, EFS, CloudWatch, ECS, SNS, SQS, Glue |
| **Databases** | MySQL, Hive, Redshift, RDS, Dynamo DB |
| **Reporting Tools/ BI tools** | Tableau, Power BI, MS Office |
| **Data Modelling Tools** | MS Visio, Lucid Chart, Relational Rose |
| **Other tools** | JIRA, GitHub, Jenkins |
| **Operating Systems** | Windows, Linux, CentOS7, Macintosh HD, Android, Unix, Mint Linux |

**PROFESSIONAL EXPERIENCE**:

**BCBS, Jacksonville, FL Mar 19 – Till Date**

**Big Data Developer**

**Responsibilities:**

* Worked on implementing pipelines and analytical workloads using big data technologies like **Hadoop, Spark, and Kafka**
* Worked with business team to understand sourcedata and performed ETL to clean and fetch the required data based on business required platforms. Proper procedure was being followed, i.e., Build **Sql script**, prepare data model document, got approval from business team and register datasets.
* Designed and developed a generic change data capture (CDC) process using **PySpark** to create the business current data.
* Handled importing data from **AWS S3 to HDFS**, performed transformation and action functions using Spark to get the desired output.
* Implemented cloud-based serverless pipelines using **AWS Lambda** to export data from Hiveinto **Redshift.**
* Designed and built **data lake** storage and **ETL** process for device insights streaming data using **Kafka** and **Spark.**
* Imported data from **S3 Glacier** to **Hive** using spark on **EMR** clusters**.**
* Developed ETL applications to execute them using **Glue**.
* Handled loading **clickstream** data from **AWS S3** to **Hive** using **crontab** and **shell scripting.**
* Created Hive UDFs for custom analytical functions used to generate business reports.
* Worked on various spark optimization techniques for **memory management, garbage collection, Serialization** and also using **Broadcast Join, Accumulators, and persisting methods**.
* Developed real time data processing applications by using **Scala and Python.**
* Closely worked with Kafka Admin team to set up Kafka cluster and implemented Kafka producer and consumer applications on Kafka cluster setup with help of **Zookeeper.**
* Developed a **Kafka Producer** application that ingests transactional data from Exadata machines to Kafka topics.
* Build a pipeline using **Spark Streaming** to receive real-time data from the **Kafka** and store the stream data to **HDFS.**
* Tune long running **Spark applications** using various optimization techniques for executor tuning, **memory management, garbage collection, Serialization**, **broadcast variable,** and **persisting methods**.
* Developed Tableau visualizations and **dashboards** using **Tableau Desktop**.
* Developed spark applications to stream data from **Kafka topics** to **HDFS** integrates with apache hive to make data immediately available for **HQL** querying.
* Developed **DDL** and **DML** scripts for creating tables and views, loading data, and develop analytics applications in **Hive** and **Redshift.**
* Implemented **python scripts** for backend database connectivity and data imports.
* Built a serverless ETL in AWS lambda to process the files that are new in the S3 bucket to be cataloged immediately.
* Built python module to access **Jira** and create issues for all the DB owners and notify them every 7 days if the issue is not closed.
* Involved in developing **CICD** pipelines using **Jenkins** to automate the code deployment.
* Involved in database design and data modeling using **Entity-Relationship modeling** for **OLTP** and **dimension modeling** for **OLAP** databases**.**
* Used **AWSSQS** to send the processed data further to the next working teams for further processing.
* Deployed **AWSLambda** functions to sync data from **MySQL** to the **Client Portal**.
* Created detailed design documentation for the source-to-target transformations and technical documentation.
* Involved in all the phases of Software Development Life Cycle (**Requirements Analysis, Design, Development, Testing, Deployment, and Support**) and **agile methodologies.**

**Environment:** Hadoop, Spark, Spark SQL, Hive, HQL, MySQL, HDFS, Scala, Shell Scripting, Crontab, Apache Kafka, Apache Spark Streaming, AWS RDS, Python, AWS Lambda, AWS EC2, JIRA, Tableau

**Wal-Mart, Bentonville, AR Jan 18 – Feb 19**

**Big Data Developer**

**Responsibilities:**

* Design and develop data pipelines for Target Integrated **Data Analysis** using **Hive, Spark, Sqoop, MySQL.**
* Involved in architecture design, development, and implementation of Hadoop deployment, backup, and recovery systems.
* Developed **MapReduce** programs in Python using Hadoop to parse the raw data, populate staging tables, and store their fined data in partitioned HIVE tables.
* Automate the **data pipelines** using **Shell Scripts, Oozie workflows,** and **Cron Jobs.**
* Involved in Designing, Development, Building Fact, and Dimension tables in **Hadoop**.
* Incorporate **GIT** for version control, **Jenkins** for Automated builds and tests, **CHEF** for Configuration management, and **OpenStack** for **Iaas.**
* Working IT partners including engineers and analysts, reporting teams, and business clients in an **Agile SCRUM** environment to give top-notch ceaselessly improving applications that are quick to deliver through a shortened **SDLC** and iterative methodology.
* Integrated **Oozie** with the rest of the **Hadoop** stack supporting several types of **Hadoop jobs** out of the box (such as **Map-Reduce, Pig, Hive, and Sqoop**) as well as system specific jobs (such as **Java programs** and **shell scripts**).
* Exploring with **Spark** improving the performance and optimization of the existing algorithms in **Hadoop** using **Spark Context, Spark-SQL, Data Frame, pair RDD’s, Spark YARN.**
* Worked in provisioning and managing multi-tenant **Hadoop clusters** on public cloud environment -**Amazon Web Services (AWS)** and on private cloud infrastructure – Openstack cloud platform.
* Expertise in working with **Hive** optimization techniques like **Partitioning, Bucketing, vectorizations** and **Map side-joins, Bucket-Map Join, skew joins,** and creating **Indexes**.
* Experience designing and executing time-driven and data-driven **Oozie** workflows.
* Developed Spark applications using spark SQL APIfor faster testing and processing of data.
* Experience in deploying data from various sources into **HDFS** and building reports using **Tableau.**
* Performed real-time analysis on the incoming data.
* Importing and exporting data into **HDFS** using **Sqoop.**
* Responsible to manage data coming from different sources.
* Perform data analysis using **Hive** and **SQL** for Adhoc and Campaign requests from business users.
* Provide functional assistance to developers for converting **SSIS** Packages into **Spark jobs.**
* Involved in converting **Hive/SQL** queries into Spark transformations using Spark **RDDs**,**Scala** and **Python.**
* Responsible to manage data coming from different sources.
* Responsible for loading data from **UNIX** file systems to **HDFS.**
* Developed **Hive UDFs** for implementing custom functions in HQL**.**

**Environment:** Spark, Spark SQL, Scala, Hive, Sqoop, Spark Streaming, HDFS, Tableau, AWS, Shell Scripting, oozie, Spark Yarn, Java, Pig, Map-Reduce, Unix, Python, GIT, Spark-context, Hive-context, IaaS, Jenkins, Agile Scrum.

**Bed Bath Beyond, Northern, NJ Oct 16 – Dec 17**

**Hadoop Developer**

**Responsibilities:**

* Experienced in designing and deploying of **Hadoop cluster** and different Big Data analytic tools including **Pig, Hive, HBase, Oozie, Sqoop, Kafka, Spark, Impala.**
* Imported weblogs and unstructured data using the **Apache Flume** and store it on HDFS.
* Loaded the **CDRs** from relational DB using **Sqoop** and other sources to Hadoop cluster by **Flume.**
* Developed business logic in **Flume** interceptor in **Java.**
* Implementing quality checks and transformations using **Flume Interceptor.**
* Worked on the **Ad hoc queries**, **Indexing, Replication, Load balancing, Aggregation** in **MongoDB.**
* Experience in managing **MongoDB** environment from availability, performance, and scalability perspectives.
* Developed multiple **MapReduce jobs** using Java for data processing.
* Developed workflows using custom **MapReduce, Pig, Hive, and Sqoop.**
* Implemented **Data Ingestion** in real-time processing using **Kafka**.
* Developed multiple **Kafka Producers** and **Consumers** as per the software requirement specifications.
* Documented the requirements including the available code which should be implemented using **Spark, Hive, HDFS.**
* Involved in converting **Hive/SQL** queries into Spark transformations using **Spark RDDs, Scala.**
* Optimized **Hive QL/ pig scripts** by using an execution engine like **Spark.**
* Worked and learned a great deal from **Amazon Web Services (AWS)** Cloud services like **EC2, S3, EBS, RDS, and VPC.**
* Worked in provisioning and managing multi-tenant Hadoop clusters on public cloud environment -**Amazon Web Services (AWS)** and on private cloud infrastructure – Openstack cloud platform
* Worked with cloud administrations like **Amazon web services (AWS).**
* Worked on POCs for using **Apache Spark** which provides the fast and general engine for large data processing integrated with functional programming language Scala.
* Experience in performance tuning a **Cassandra cluster** to optimize it for writes and reads.
* Experience in Data modeling and connecting **Cassandra** from spark and saving the summarized data frame to **Cassandra.**
* Experience designing and executing time-driven and data-driven **Oozie** workflows.

**Environment:** Hadoop, HDFS, MapReduce, Pig, Hive, pig scripts, YARN, Sqoop, Oozie, Python, AWS, Shell Scripting, Impala, Spark, spark-SQL, Talend, Cassandra, SOLR, Zookeeper, Scala, Kafka, Cloudera, MongoDB.

**Bigtec Pvt Ltd, India Jul 14 – Sep 16**

**Java Developer**

**Responsibilities:**

* Full life-cycle development and project management of Fee and Billing application, starting from problem identification and leading through requirements gathering, design, coding, testing, documentation, and delivery.
* Preparation of Technical architecture and Low-level design documents
* Coordinate with the team to deliver the modules on the target
* Involved in Unit and performance testing the system
* Verifying the quality of the requirements, including testability, requirement definition, test design, test-script and test-data development, test automation, test-environment configuration, test-script configuration management, and test execution.
* Used Servlets to implement Business components.
* Designed and Developed required Manager Classes for database operations
* Developed various Servlets for monitoring the application.
* Designed and developed the front-end using HTML and JSP
* Developed XML files, DTDs, Schema's, and parsing XML by using both SAX and DOM parser.
* Wrote deployment descriptors using XML and Test java classes for direct testing of the Session and Entity beans.
* Did Packaging and Deployment of builds through ANT script.
* Wrote stored procedures and used JAVA APIs to call these procedures.
* Database designing includes defining tables, views, constraints, triggers, index, stored procedures, etc.
* Developed verification and validation scripts in java.
* Preparation of Traceability Matrix to ensure use case coverage
* Perform Design and Code Review and signoff of Design document and Test Plan

**Environment:** Java 1.5, Struts, Spring, Hibernate, Eclipse, and SQL Server 2005, ANT, XML

**Avaya, India June 12 – Jun 14**

**Java Developer**

**Responsibilities:**

* Involved in eliciting the requirement, use case modeling, design, leading the development team, and tracking the project status.
* Understanding the Functional and Technical Requirement Specification of the customer’s need for the product.
* Impact analysis of the application under development.
* Preparing Design Specification, User Test Plans, and Approach Notes for the application under development.
* Creating UML Diagrams like Class Diagram, Sequence Diagram, and Activity Diagram.
* Developing application as per given specifications.
* Creation of User-based modules implementing different levels of security.
* Involved in designing of Front-end, Implementing Functionality with Business Logic.
* Developed java programs, JSP pages, and servlets using the Cantata Struts framework.
* Involved in creating database tables, writing complex TSQL queries, and stored procedures in the SQL server.
* Worked with the AJAX framework to get the asynchronous response for user requests and JavaScript for the validation.
* Used EJBs in the application and developed Session beans to implement business logic at the middle tier level.
* Actively involved in writing SQL using SQL Query Builder.
* Used JAXB to read and manipulate the XML properties.
* Used JNI for calling the libraries and other implemented functions in C language.
* Handling Server Related issues, new requirement handling, changes, and patches movements.
* Mentoring and grooming juniors technically on Agile practices and Java/J2EE development issues.

**Environment:** Java, Servlets, JSP, EJB, XML, JUnit and Eclipse, Oracle, SQL, UNIX, UML

**Education:**

* Bachelor’s in Computer Science and Engineering, INDIA in 2012