**Vikas J**

VikasJ0203@gmail.com

+1 (940-784-3894)

Seattle WA

**Professional Summary:**

* Around 8 years of experience in IT and 6+ years of specialized knowledge in BIGDATA using HADOOP and its related tools.
* In-depth understanding of Hadoop technologies such as MapReduce, Pig, Oozie, Hive, Zookeeper, Sqoop, Scala, HBase, Hortonworks, Hue, and Cloudera.
* Proven expertise in designing and implementing complete big data solutions with Hadoop tools like HDFS, MapReduce, Hive, Pig, Spark, Oozie, and Flume.
* Skilled in transferring data from various sources like RDBMS, NoSQL databases, flat files, and log files into Hadoop using Sqoop and Flume
* Experience in utilizing Hive and Pig for data warehousing and modeling within Hadoop.
* Hands-on experience in Spark for batch processing, real-time streaming, and SQL queries.
* In-depth knowledge of HBase architecture, data modeling, and optimization techniques.
* Expertise in designing and deploying Oozie workflows for scheduling Hadoop jobs.
* In-depth understanding of Hadoop cluster administration and management using tools such as Cloudera Manager and Hortonworks Ambari.
* Familiarity with Hadoop security implementation using Kerberos and Sentry.
* Ability to solve problems, troubleshoot issues, and optimize Hadoop job performance.
* Proficient in Java and Scala programming for custom MapReduce development.
* Strong knowledge of Linux administration, shell scripting, and programming languages such as Python.
* 2+ years of extensive experience in designing and developing Tableau dashboards and visualizations with Tableau Desktop and Server Administration.
* 4+ years of hands-on experience in utilizing Business Objects products such as Web Intelligence, Desktop Intelligence, Info Universe Designer, View, CMC, CCM, and CMS.
* A proven record of working as a Data Engineer for cloud services, Big Data/Hadoop applications, and product development projects.
* Skills in creating interactive dashboards with Tableau that meet the needs of the business.
* In-depth knowledge of managing and maintaining Tableau Server, including security settings, report scheduling and publishing, and user access control.
* Expertise in Business Objects universe design, report creation, and web intelligence reporting.
* Proficient in Amazon Web Services (AWS) and experience in cloud-based big data solutions
* Strong ability to collaborate with cross-functional teams and communicate data insights to stakeholders.
* Track record of designing and implementing data solutions that fulfill business requirements and enhance decision-making processes.
* Familiarity with data warehousing concepts, data modeling, and data integration techniques.
* Ability to create meaningful visualizations and dashboards that help stakeholders make informed decisions.
* A proven track record of working as a Data Engineer on Amazon cloud services, big data/Hadoop applications, and product development.
* Strong knowledge of Informatica & DataStage administration, production deployment, and working with IBM/Informatica.
* Skills in data warehousing concepts, dimensional modeling, data modeling, OLAP, and OLTP systems.
* Involved in the key role of migrating Cassandra and Hadoop clusters to AWS and defined various read/write strategies.
* Expertise in SQL development, including writing stored procedures, triggers, views, and user-defined functions.
* Proficient in using Informatica Power Center 9.x/8.x for ETL (Extract, Transform, Load) processes.
* Proficient in multiple databases such as MongoDB, Cassandra, MySQL, Oracle, and MS SQL Server.
* Proficient in programming languages like Java, Python, R, and SAS.
* Experienced in using various components of the Hadoop ecosystem like HDFS, YARN, MapReduce, Spark, Pig, Sqoop, Hive, Impala, HBase, Kafka, and Crontab tools.
* Capable of creating HIVE UDFs in Java to analyze complex data sets.
* Experience in developing ETL applications on large data sets using MapReduce, Spark-Scala, PySpark, Spark-SQL, and Pig.
* Experienced in using SQOOP to import and export data between RDBMS and HDFS/Hive.
* Strong knowledge of MS SQL Server, including SSRS, SSIS, and T-SQL.

# SKILLS:

**Hadoop Ecosystem:** HDFS, SQL, YARN, PIG Latin, MapReduce, Hive, Sqoop, Spark, Yarn, Zookeeper, Oozie, Kafka, Storm, Flume.

**Programming Languages:** Python, Scala, Java, PySpark, JavaScript, Shell Scripting.

**Big Data Platforms:** Hortonworks, Cloudera.

**Cloud Systems:** Kubernetes, Docker, Amazon Web Services, Azure, Google Cloud Platform Services.

**Amazon Web Services:** EC2, S3, EMR, Redshift, DynamoDB, Aurora, VPS, Glue, Kinesis, Boto3

**Google Cloud Platform Services**: Cloud Storage, Cloud DataProc, Cloud SQL, Cloud Functions, Composer, Big Query, Cloud Pub/Sub.

**Microsoft Azure**: Active Directory, Applications Insights, Azure Monitoring, Key Vault, Blob Storage, Azure Search, Data Factory, Azure Databricks, SQL Azure, and Azure Analysis Services.

**Operating System:** Windows, DOS, Unix, Linux.

**Databases:** Netezza, MySQL, UDB, HBase, MongoDB, Cassandra, Snowflake.

**Development Methods:** Agile/Scrum, Waterfall.

**IDE’s:** PyCharm, IntelliJ, VsCode Jupyter Notebook.

**Data Visualization:** Tableau, BO Reports, Splunk.

**PROJECTS:**

**Software Developer Engineer-I / Data Platform June 2022 – Present**

**Amazon (AWS SageMaker MLF), Seattle, WA**

**Responsibilities:**

* Designed and Developed multiple ETL data pipelines to Redshift for financial data of terra byte scale using AWS tools such as S3, EMR, IAM, and EC2.
* Represented my team in the migration of reports built in OBIEE (Oracle Business Intelligence Suite Enterprise Edition) to Quicksight. Evaluated 70 reports and formulated the plan for migration for my team in 3 months, we developed, tested, and delivered 50 critical reports in Quicksight used by various teams for financial reporting.
* Built interactive dashboards in AWS QuickSight
* Productized Data Export APIs to allow data download from Data Lake with ad-hoc queries.
* Automation implementations using Python, Unix shell scripting excel macros.
* Hands-on experience in building, designing, and maintaining cloud-based applications with AWS CDK.
* Worked on Open-Source Public GitHub Repositories (aws/sagemaker-python-sdk and aws/sagemaker-spark).
* Extensive experience in Build/Release Management GitHub along with Shell Scripting.
* Understanding the Requirement Specifications and Design Documents of the application.
* Wrote Python Scripts to manage AWS resources from API calls using BOTO SDK and worked on AWS CLI.
* Worked on Amazon SageMaker Studio, Studio Lab, and Notebook.
* Utilized AWS services such as HEX-AWS to store and process large amounts of data.
* Used AWS Glue for data transformation, validation, and cleansing.
* Utilized Python Boto 3 to configure AWS glue, EC2, and S3 services.
* Automated data transfer using Python scripts, involved in file processing, splitting, and validation.

**Tools and Technologies**: Python, Spark, Hadoop, Scala, Aws Services, SageMaker, EMR, IAM, EC2, S3, Glacier, DynamoDB, Redshift, Oracle DB, AWS Lambda, CloudWatch, PyTest, IntelliJ, Visual Studio, Dockers, and HEX-AWS

**Centene Corporation, St Louis May 2021 – May 2022**

**Data Engineer**

**Responsibilities:**

* Led migration of “1” on-premises data warehouses to Azure Synapse Analytics, enhancing retrieval times by 70 %.
* Designed and implemented real-time analytics solutions using Azure Stream Analytics and Power BI, providing actionable insights for business teams.
* Developed automated data pipelines using Azure Data Factory, integrating data from various sources including on-prem databases, web services, and third-party cloud platforms.
* Collaborated with data science teams, optimizing Azure Databricks operations for ML model training.
* Designed and implemented real-time data streaming solutions using Azure Event Hubs and Azure Stream Analytics, enabling the business to react promptly to live data feeds.
* Constructed and maintained data models in Azure Synapse Analytics to support BI tools, ensuring fast and efficient data retrieval for end-users.
* Leveraged Azure Functions to create serverless data processing solutions, reducing infrastructure costs and streamlining deployment.
* Set up automated data quality checks and monitoring within Azure, ensuring the accuracy and reliability of analytical reports.
* Designed and executed disaster recovery plans for Azure data services, ensuring minimal downtime and data loss during unforeseen events.

**Environment:** Azure, Python, Spark, ETL, My SQL, Oracle, Linux, HDFS, Tableau Desktop 9/10, Linux, Data visualization in D3, Microsoft Excel.

**Data Engineer June 2017 – Dec 2020**

**Broadridge Financial Solutions - Hyderabad, India**

**Responsibilities:**

* Orchestrated the setup and configuration of Azure Databricks workspaces, clusters, and jobs, ensuring optimized performance and cost management.
* Developed robust ETL pipelines using Databricks to process, clean, and transfer multi-terabyte datasets to and from Azure Data Lake Storage, improving data processing speeds by 50%.
* Leveraged Delta Lake's capabilities to build a high-performance, reliable, and simplified data lake, ensuring ACID compliance and efficient upsert operations.
* Implemented structured streaming in Databricks for real-time analytics, enabling actionable insights from live data feeds.
* Conducted performance tuning of Spark jobs in Databricks, utilizing best practices like broadcasting, caching, and partitioning, achieving a 15% reduction in processing times.
* Used Databricks' integrated monitoring tools and integrated with Azure Monitor and Log Analytics to track job performance, failures, and resource utilization.
* Implemented auto-scaling and termination policies for Databricks clusters, optimizing resource usage and reducing Azure costs by 20%.
* Transformed raw data into a more efficient and optimized data format such as Avro and Parquet to speed up data processing and improve data transfer through networks.
* Conducted data analysis, auditing, forecasting, programming, research, report generation, and software integration to gain an in-depth understanding of the current BI platform architecture and support the implemented solution.
* Designed, implemented, and tested various performance tests to ensure the successful delivery of a project. Utilized data-driven performance analytics to produce reports and dashboards with valuable insights.
* Designed, developed, and tested ETL applications with different sources.
* Constructed and optimized SQL queries in Hive through HUE. Executed MapReduce jobs in Hive by querying available data.

**Environment:** Python, Spark, Azure Databricks, Tableau (9.x/10.x), HDFS, MapReduce, Flume, Kafka, Zookeeper, Pig, Hive, HQL, HBase, Spark, Kafka, PostgreSQL, AWS, JIRA, GitHub.

**Data Analyst/Engineer Sep 2015 – April 2017**

**ITCELL INC - Hyderabad**

**Responsibilities:**

* Designed, built, and managed Big Query data warehousing solutions to handle multi-terabyte datasets, achieving a 30 % reduction in query times and costs.
* Developed and deployed scalable stream and batch data processing pipelines using Dataflow and Apache Beam, processing over 4 TBs of data monthly.
* Automated data ingestion and synchronization tasks using Cloud Data Transfer and Transfer Service for on-premises data.
* Orchestrated the secure storage and management of data in Google Cloud Storage, implementing lifecycle policies, and versioning to optimize storage costs.
* Provisioned and managed Hadoop clusters using Cloud Dataproc, optimizing Spark and Hadoop jobs to reduce processing times by 20%.
* Collaborated with business analysts to visualize and report data insights using GCP Data Studio, enhancing decision-making processes.
* Responsible for identifying, gathering, exploring, and cleansing data for modeling.
* Utilized HIVE to execute transformations, events, and some pre-aggregations before storing the data in HDFS.
* Authored SQL and PL/SQL scripts to extract data from databases and for testing purposes.
* Designed a Spark batch job to automate the creation/update of external Hive tables from datasets stored in HDFS.
* Worked with HDFS file formats such as Avro, Sequence File, and various compression formats such as Snappy.
* Created Spark SQL scripts to optimize query performance.
* Implemented Spark scripts using Spark Session, Python, and Spark SQL to access Hive tables for more efficient data processing.
* Partnered with data engineers and operation team to implement ETL processes, Snowflake models, and optimized SQL queries to extract data to meet analytical requirements.
* Applied Dimensional Data Modeling to deliver Multi-Dimensional STAR schemas and developed Snowflake schemas by normalizing dimension tables as needed.
* Revamped ETL procedures, added data management and data query components, and updated existing ones.
* Employed Redshift and S3 within AWS along with Informatica Cloud Services to load data into an S3 bucket.
* Conducted data analysis and visualizations using Python and Tableau.

**Environment:** Python, GCP, HDFS, AWS (S3, Redshift), MapReduce, Kafka, Hive, HBase, Spark, Kafka, ETL, Web Services, Linux RedHat, SQL Server.

**EDUCATION AND TRAINING**

**Bachelor of Technology – Electronics and Instrumentation Engineering**

Keshav Memorial Institute of Technology, May 2016

**Masters – Data Science**

University of North Texas, May 2022.